http://dx.doi.org/10.18778/7525-695-6

Jerzy Korzeniewski

Metody selekcji zmiennych w analizie skupień

Nowe procedury

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

ŁÓDŹ 2012

RECENZENT
Grażyna Trzpiot

PROJEKT OKŁADKI
Barbara Grzejszczak

Wydrukowano z gotowych materiałów dostarczonych do Wydawnictwa UŁ

© Copyright by Uniwersytet Łódzki – Wydawnictwo Uniwersytetu Łódzkiego 2012

ISBN 978-83-7525-695-6

Wydawnictwo Uniwersytetu Łódzkiego
90-131 Łódź, ul. Lindleya 8

Wydanie I. Nakład 141 egz. Ark. druk. 11,875
Zam. 5050/2012. Cena zł 26,– + VAT

SPIS TREŚCI

1. Wprowadzenie 5
1.1. Cele pracy 5
1.2. Charakterystyka analizy skupień. Wybór mierników 7
1.3. Uwarunkowania procesu wybierania zmiennych w analizie skupień 25
1.4. Zbiory danych generowanych 34
1.4.1 Zbiory danych ze zmiennymi ciągłymi 34
1.4.2 Zbiory danych ze zmiennymi porządkowymi 39
1.4.3 Zbiory danych ze zmiennymi binarnymi 42

2. Metody modelowe wybierania zmiennych w analizie skupień 44
2.1. Uwagi wstępne 44
2.2. Metoda wyrazistości zmiennych 45
2.3. Metoda wyboru modelu 46
2.4. Metoda separowalności rozrzutu 48
2.5. Ocena podejść modelowych 49

3. Metody heurystyczne wybierania zmiennych w analizie skupień. 51
3.1. Uwagi wstępne 51
3.2. Metoda Brusco dla zmiennych binarnych 56
3.3. Metoda Talavery-Fishera dla zmiennych nominalnych. 59
3.4. Metoda COSA 61
3.5. Metoda kolejnych rzutowań 64
3.6. Metoda HINoV 65
3.7. Metoda VS-KM 69
3.8. Metoda VAF z ważeniem zmiennych indeksem skupialności 72
3.9. Metody oparte na entropii 75
3.10. Metoda uśredniania zmiennych 78
3.11. Metody oparte na grupowaniu rzadkim 80
3.12. Metoda Ichino oparta na teorii grafów 82

4. Nowa metoda selekcji zmiennych w analizie skupień 86
4.1. Korelacja odległościowa i jej własności 86
4.2. Wymuszanie wyższej wartości korelacji odległościowej 100
4.3. Kilkustopniowe grupowanie obiektów w dwa skupienia 107
4.4. Sformułowanie metody dla różnych rodzajów danych 114
4.4.1. Metoda dla zbiorów danych ze zmiennymi ciągłymi 114
4.4.2. Metoda dla zbiorów danych ze zmiennymi porządkowymi 119
4.4.3. Metoda dla zbiorów danych ze zmiennymi binarnymi 124
4.5. Ocena nowej metody przy pomocy eksperymentów symulacyjnych 127

4.5.1. Eksperyment dla zbiorów danych ze zmiennymi ciągłymi 127
4.5.2. Eksperyment dla zbiorów danych ze zmiennymi porządkowymi ..131
4.5.3. Eksperyment dla zbiorów danych ze zmiennymi binarnymi 132
4.6. Sformułowanie metody dla dowolnego układu skal pomiarowych 133

5. Zastosowanie metod selekcji zmiennych w analizie skupień w badaniach ekonomiczno-społecznych.....139
5.1. Zagadnienia wstępne dotyczące empirycznych zbiorów danych 139
5.2. Metodologia oceny dla empirycznych zbiorów danych 142
5.3. Badanie efektywności metod na empirycznych zbiorach danych 143
5.4. Ocena efektywności aplikacyjnej metod selekcji zmiennych 175
Zakończenie 176
Literatura.....179
Załącznik 1. Oznaczenia i symbole.....186
Załącznik 2. Spis programów komputerowych.....188
Od redakcji.....189
Summary.....190

1. Wprowadzenie

1.1. Cele pracy

Analiza skupień (*cluster analysis*) zwana też nauczaniem bez nadzoru (*unsupervised learning*) lub taksonomią (*taxonomy*) poświęcona jest zagadnieniu pogrupowania obiektów zbioru danych w rozłączne grupy (skupienia) spójne wewnętrznie i jak najbardziej zróżnicowane pomiędzy sobą. Poprawnie przeprowadzona analiza skupień umożliwia podzielenie zbioru danych na grupy, celem lepszego zrozumienia informacji w nich zawartych i ustalenia własności grup obiektów do siebie podobnych oraz ich syntetycznej charakterystyki. Analiza skupień odgrywa bardzo ważną rolę w różnych dziedzinach nauki, w tym w analizach zjawisk społeczno-ekonomicznych. Typowym przykładem może być segregacja klientów banku lub firmy na grupy klientów podobnych do siebie pod względem cech istotnych dla banku lub firmy. Analiza skupień rozwija się od ponad 50 lat. W tym czasie opracowano wiele różnych metod realizujących zadania cząstkowe analizy skupień, które nazywane są dzisiaj etapami analizy skupień (por. § 1.2). W ostatnich latach, wzorem analogicznej drogi w przypadku klasyfikacji (nauczania z nadzorem) rozwija się podejście zagregowane do analizy skupień (*ensemble clustering*), którego ideą jest łączenie wyników uzyskanych przy pomocy różnych metod, mające na celu poprawienie wyników. Realizacja poszczególnych etapów analizy skupień musi być wykonywana w oparciu o zbiór zmiennych charakteryzujących obserwacje. W zagadnieniach praktycznych zawsze stoimy przed koniecznością określenia i wyboru zbioru zmiennych, którymi będziemy się posługiwać. Jeśli zmienne zostaną niewłaściwie wybrane, to fakt ten może w znacznym stopniu obniżyć jakość otrzymanych wyników. Jak zaznacza się w literaturze wybór zmiennych w klasyfikacji jest jednym z najważniejszych a zarazem najtrudniejszych zadań. Zmienne, które tworzą strukturę skupień w zbiorze danych nazywać będziemy istotnymi (*relevant features, true features*) zaś pozostałe zmienne nieistotnymi lub zakłócającymi (*irrelevant features, masking features, noisy features).*

Cele rozprawy można ująć następująco:

1. Ocena dotychczasowego dorobku naukowego w zakresie wybierania zmiennych tworzących strukturę skupień w zbiorze danych;
2. Zaproponowanie nowych rozwiązań, które będą efektywniejsze od metod istniejących, nieobarczone tyloma założeniami co najlepsze dotychczas skonstruowane metody, założeniami takimi jak:

- konieczność znajomości liczby skupień w zbiorze danych;
- ograniczenia możliwości stosowania w przypadku występowania zmiennych mierzonych na różnych skalach;
- ograniczenie do wyboru o charakterze porównawczym pomiędzy dwoma zbiorami zmiennych, czego konsekwencją jest brak możliwości wyszukiwania wielokrotnych struktur danych;
- konieczność odwołania się do subiektywnej oceny stosowanych wskaźników.

3. Zbadanie efektywności nowych rozwiązań w zastosowaniu ich do selekcji zmiennych w empirycznych zbiorach danych o charakterze ekonomiczno-socjologicznym.

Pierwszy rozdział pracy jest poświęcony zagadnieniom wstępnym i zawiera charakterystykę etapów analizy skupień, ogólne omówienie uwarunkowań zagadnienia wybierania zmiennych w analizie skupień oraz stosowane w dalszym ciągu rozprawy charakterystyki pojedynczych zmiennych, miary podobieństwa podziałów zbioru danych, miary korelacji zmiennych, miary jakości wyników dokonanego wyboru zmiennych. Ponadto, przedstawione są zbiory danych generowanych stosowane później w eksperymentach symulacyjnych.

Zawartość drugiego oraz trzeciego rozdziału podzielona została według kryterium odnoszącego się do tego czy metoda oparta jest na wnioskowaniu statystycznym na podstawie modelu czy też jest algorytmem czysto heurystycznym nie opartym na modelu. Alternatywnym podziałem mógłby być, na przykład, podział na metody dokonujące selekcji zmiennych i metody ważące zmienne. Autor niniejszej monografii stoi na stanowisku, że taki podział byłby o wiele mniej różnicujący, gdyż selekcja zmiennych jest szczególnym przypadkiem ważenia zmiennych. Ponadto, bardzo istotne jest też to, że problematyka metod modelowych jest podobna dla wszystkich tego typu metod, gdyż problemem zasadniczym jest estymacja parametrów modelu, na ogół, dokonywana przy wykorzystaniu tego samego algorytmu. Inną alternatywą mógłby być, na przykład, podział metod ze względu na rodzaj skali pomiarowej zmiennych opisujących obiekty zbioru. Ta alternatywa wydaje się mniej rozsądna ze względu na to, że niektóre metody można stosować zarówno do zbiorów danych z silnymi jak i słabymi skalami pomiarowymi.

Czwarty rozdział zawiera propozycję autorską nowych metod wyboru zmiennych w analizie skupień oraz wyniki badania tej metody za pomocą eksperymentów Monte Carlo. Ideą nowej metody jest zastąpienie korelacji pomiędzy zmiennymi opisującymi obserwacje korelacją pomiędzy odległościami par obiektów oraz zastąpienie grupowania na określoną liczbę skupień grupowaniem wielokrotnym na dwa skupienia. Uzyskujemy dzięki temu

uniwersalną metodę, którą możemy stosować zarówno do słabych jak i silnych skal pomiarowych mierzących wartości zmiennych. Ponadto, do pewnego stopnia uwalniamy się od konieczności znajomości liczby skupień w zbiorze danych.

Piąty rozdział poświęcony jest stronie aplikacyjnej zagadnienia tj. ocenie porównawczej zaproponowanych nowych metod oraz najlepszych dotychczas skonstruowanych metod na podstawie efektywności uzyskanej w zastosowaniu do empirycznych zbiorów danych. Do tej pory nie udało się skonstruować metody nieobarczonej co najmniej kilkoma założeniami i takiej, która dobrze wybierałaby zmienne dla zbiorów danych opisanych przez zmienne mierzone na różnych skalach i dawała dobre efekty zarówno dla dużych (w sensie liczby zmiennych) jak i dla małych zbiorów danych. Próba podjęta przez autora ma na celu znalezienie takiej metody, która dla badaczy będzie najbardziej wiarygodna – odporna na niektóre zakłócenia.

Problem selekcji zmiennych w analizie skupień można rozpatrywać w różnych ujęciach w odniesieniu do założeń mniej lub bardziej ograniczających stosowalność konstruowanych metod. Niektóre ograniczenia zostały wymienione w celach rozprawy. Należy jeszcze wspomnieć o tym, że w swej najogólniejszej postaci, w problemie selekcji dopuszcza się możliwość istnienia kilku struktur skupień w tym samym zbiorze danych, przy czym zbiory zmiennych tworzących te struktury nie muszą być rozłączne. Ta najogólniejsza forma problemu jest na tyle skomplikowana, że metody, które ją rozważają spisują się bardzo słabo w zawężonych formach problemu, na przykład, w najprostszej postaci, w której zakładamy, że istnieje tylko jeden zbiór zmiennych tworzących strukturę skupień. Propozycje autora mają przede wszystkim na uwadze ujęcie problemu, w którym zakładamy rozłączność zbiorów zmiennych tworzących różne struktury skupień.

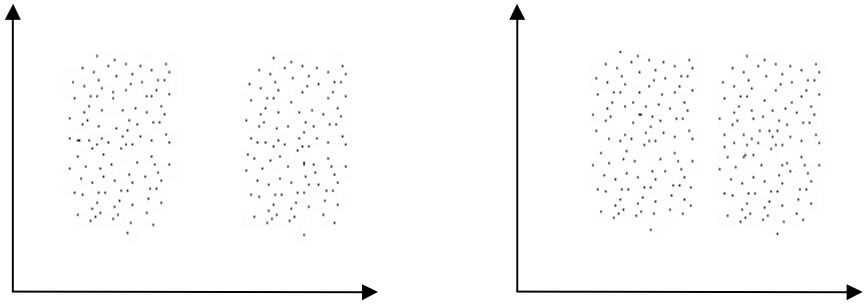
W zakończeniu przedstawione są wnioski z przeprowadzonych badań podsumowujące wszystkie uzyskane oceny metod istniejących oraz propozycji autora.

Rozprawa niniejsza była wspomagana środkami z grantu habilitacyjnego nr 4323/B/H03/2009/37.

1.2. Charakterystyka analizy skupień. Wybór mierników

Analiza skupień jest dziedziną statystyki bardzo istotną pod względem aplikacyjnym. Konsekwencją potrzeby opracowania metod niezbędnych do segmentacji zbiorów danych na spójne podgrupy obiektów przy braku jednoznacznej i precyzyjnej definicji takich podgrup jest pewien chaos w dorobku tej dziedziny. Niewiele jest wyników teoretycznych, na przykład, o wiele mniej niż w pokrewnej dziedzinie statystyki jaką jest klasyfikacja ze

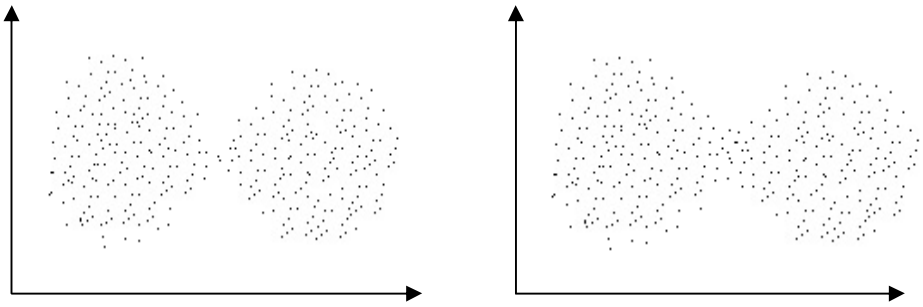
zbiorem uczącym, natomiast bardzo wiele różnych algorytmów i metod heurystycznych.



Rys.1.1. Dwa przykłady zbiorów punktów dwuwymiarowych, z których jeden tworzy bardzo wyraźną strukturę dwóch skupień.

Ź r ó d ł o: opracowanie własne.

Przyczyną takiego stanu rzeczy jest to, że nie ma jednoznacznej definicji skupienia i, co za tym idzie, często trudno zdecydować czy w danym zbiorze istnieje struktura skupień czy też nie. Na rysunkach 1.1 oraz 1.2 przedstawione są przykłady czterech zbiorów punktów dwuwymiarowych. Zbiory po lewej stronie tworzą wyraźną lub bardzo wyraźną strukturę skupień. Zbiór po prawej stronie składa się z podobnych podzbiorów punktów ale orzeczenie o istnieniu struktury skupień zależy od wizualnych upodobań patrzącego.



Rys.1.2. Dwa przykłady zbiorów punktów dwuwymiarowych, z których jeden tworzy dość wyraźną strukturę dwóch skupień.

Ź r ó d ł o: opracowanie własne.

Powyższe, dość lapidarne sformułowania, precyzyjniej ujął światowej klasy ekspert w dziedzinie informatyki, sztucznej inteligencji i możliwości

wykorzystania komputerów do rozwiązywania problemów świata realnego, Jon Kleinberg. Wyróżnił on (*Kleinberg,* 2002) trzy własności, które w analizie skupień byłyby bardzo pożyteczne. Pierwsza własność to niezmienniczość grupowania względem skali odległości (*scale invariance*) polegająca na tym, że dla dowolnego skalara c i dowolnej miary odległości d funkcja f decydująca o grupowaniu obiektów powinna mieć własność, którą symbolicznie można ująć jako $f(d) = f(c \cdot d)$. Druga własność to zupełność grupowania (*richness*) polegająca na tym, że dla jakiegokolwiek podziału P powinna istnieć odległość d taka, że $f(d) = P$. Trzecia własność to zgodność grupowania (*consistency*) polegająca na tym, że $f(d) = P$ implikuje $f(d^*) = P$, gdzie odległość d^* jest transformacją odległości d indukowaną przez podział P w następujący sposób: odległość indukowana dowolnych dwóch obiektów z tego samego skupienia w podziale P nie przekracza odległości d tych obiektów oraz odległość indukowana dowolnych dwóch obiektów z różnych skupień w podziale P jest co najmniej taka jak odległości d tych obiektów. Następnie Kleinberg udowodnił twierdzenie o niemożności (*impossibility theorem*): nie ma funkcji decydującej o grupowaniu obiektów, która miałaby wszystkie trzy własności. Można polemizować z tym czy sformułowane własności są rzeczywiście niezbędne w każdej aplikacji empirycznej analizy skupień, ale, z pewnością, twierdzenie o niemożności rozwiewa wiele wątpliwości nękających wszystkich badaczy starających się znaleźć jak najlepszą metodę do grupowania obiektów badanego zbioru danych.

W czasie kilkudziesięciu lat rozwoju analizy skupień wypracowano zgodę co do tego, że pełna analiza skupień zbioru danych powinna objąć następujące etapy:

1. Wybór obiektów i zmiennych.
2. Wizualizacja obiektów (lub zmiennych).
3. Normalizacja zmiennych.
4. Wybór miary odległości pomiędzy obiektami.
5. Wybór metody grupowania obiektów.
6. Ustalenie liczby skupień.
7. Grupowanie obiektów – właściwy etap analizy skupień.
8. Ocena wyników grupowania.
9. Opis i profilowanie klas.

Sokołowski (1992) ujmuje zadania analizy skupień jeszcze ogólniej, włączając do charakterystyk zbioru czynnik czasowy. Następnie dzieli on zadania taksonomiczne na proste (np. grupowanie obiektów), złożone (grupowanie obiektów i cech) i kompleksowe.

Wybór obiektów zbioru danych polega na ustaleniu czy istnieją w zbiorze jakieś obserwacje przypadkowe będące, na przykład konsekwencją błędów pomiarowych lub sposobu zbierania danych, które nie powinny być uwzględniane w dalszej analizie. Taka procedura łączy się z pojęciem odporności i, jak wiadomo, jest ściśle powiązana z wyborem cech opisujących obiekty a nawet doborem metod do końcowych etapów analizy skupień, na przykład, metod grupowania obiektów. Może się bowiem okazać, że dla niektórych metod grupowania, uwzględnienie wszystkich obiektów prowadzi do mniejszych strat jakości niż usuwanie obiektów podejrzanych o to, że są obiektami nietypowymi (*outliers*). Ponadto wybór obiektów jest ściśle związany z wyborem zmiennych opisujących obiekty. Selekcja zmiennych jest jednym z najważniejszych i jednocześnie najtrudniejszych etapów analizy skupień. Uwarunkowania tego etapu zostały przedstawione bardziej szczegółowo w następnym paragrafie.

Etap wizualizacji danych jest ważny dla analizy skupień, gdyż jest pomocny w odkryciu ewentualnej struktury skupień w zbiorze danych, ich liczby, a nawet przy wyborze właściwych algorytmów grupowania obiektów (biorąc pod uwagę ich własności). Wizualizacji danych można dokonać nie tylko dla wartości oryginalnych zmiennych, lecz również dla macierzy odległości między zmiennymi. W tym zakresie, oprócz tradycyjnych wykresów dwuwymiarowych, pomocne są (por. np. *Everitt i inni,* 2001): wielowymiarowe wykresy rozrzutu; trójwymiarowe wykresy zmiennych; metody skalowania wielowymiarowego i sieci Kohonena, umożliwiające graficzne przedstawienie danych w przestrzeni o mniejszej liczbie wymiarów. Program R pozwala dodatkowo tworzyć bardziej złożone wykresy (*Gatnar, Walesiak,* 2009): wielowymiarowy wykres rozrzutu z funkcją gęstości poszczególnych zmiennych; wykres rozrzutu dla 3 zmiennych metrycznych w przestrzeni 2-wymiarowej *(bubbleplot)*. Metody graficzne można nawet, w przypadkach prostych zbiorów danych, wykorzystać we wcześniejszym etapie analizy skupień tj. etapie doboru zmiennych, np. wykres pudełkowy (ramka-wąsy, *box-whiskers*).

Ewentualna normalizacja zmiennych jest traktowana jak szczególny przypadek ważenia zmiennych, gdyż standaryzacja (lub inne przekształcenie cech) wpływa na rozmieszczenie obiektów w przestrzeni euklidesowej. Już Cormack (1971) zauważył ten problem i doszedł do wniosku, że na ogół standaryzacja zmiennych zmniejsza efektywność analizy skupień. Cel standaryzacji czyli wyeliminowanie różnic w skalach pomiaru cech, jest niespójny z zasadniczym celem analizy skupień, gdyż różnice pomiędzy cechami mogą wynikać z ich naturalnych własności implikujących istnienie ewentualnej struktury skupień. Dowiedziono, że nadawanie zmiennym wag odwrotnie proporcjonalnych do całkowitej zmienności cech – czyli standaryzacja zmiennych przez ich całkowite odchylenie – jest nieefektywny, a

wręcz niewskazany, gdyż utrudnia rozróżnianie grup obiektów podobnych. Tezę o tym, że normalizowanie pojedynczych zmiennych może mieć negatywny wpływ na zachowanie oryginalnej struktury skupień tzn., może tę strukturę zniekształcić lub nawet zniszczyć stawia również Stoddard (1979).

Milligan (1996) wskazuje na błędne przekonanie wielu badaczy, że fakt występowania znacznej różnicy zmienności cech w analizie skupień jest podstawą do przeprowadzenia standaryzacji zmiennych z obawy o to, by cechy o dużej zmienności nie miały nadmiernego wpływu na wyniki analizy skupień. Podkreśla, że normalizacja zmiennych jest kwestią indywidualną, a nie rutynowym przekształceniem jak również, że nie zawsze uzasadnione jest twierdzenie, że standaryzacja może ukryć ewentualną strukturę skupień występującą w zbiorze danych. Powołując się na wcześniejsze badania porównawcze, Milligan rozważa również ewentualne procedury normalizacji zmiennych. Z reguły przyjmuje się klasyczną normalizację zmiennych, jeśli natomiast rozważymy separowalność i wewnątrzgrupową zmienność skupień otrzymanych na podstawie cech unormowanych, to okazuje się, że na tle innych przekształceń jest to rozwiązanie nieefektywne. Zbadano osiem sposobów normalizacji (*Milligan, Cooper*, 1988), (0) brak normalizacji, (1) normalizacja za pomocą klasycznej standaryzacji, (2) standaryzacja przez odchylenie standardowe, (3) przekształcenie ilorazowe w oparciu o wartość maksymalną, (4) unitaryzacja przez rozstęp zmiennej, (5) unitaryzacja zerowana, (6) przekształcenie ilorazowe w oparciu o sumę wariantów cechy i (7) rangowanie. Najlepsze wyniki pod względem stopnia zgodności klasyfikacji z właściwą strukturą skupień (ocena indeksem Randa) za pomocą metod aglomeracyjnych uzyskano dla przekształceń normalizacyjnych opartych na rozstępie zmiennych (typ (4) i (5)).

Zasadność takiej tezy dotyczącej ważenia (normalizacji) zmiennych na potrzeby analizy skupień potwierdzają wyniki badań, które przedstawili Gnanadesikan i inni (1995). Autorzy zastosowali dziewięć procedur ważenia cech, m.in.: (0) brak ważenia, (1) standaryzację zmiennych z wykorzystaniem samego odchylenia standardowego, (2) unitaryzację opartą tylko na rozstępie, (3) macierz odwrotną wewnątrzklasowej zmienności cech, (4) macierz odwrotną elementów diagonalnych macierzy wewnątrzklasowej zmienności cech, (5) iloczyn elementów diagonalnych macierzy międzyklasowej i elementów diagonalnych macierzy odwrotnej do wewnątrzklasowej, spośród których najlepsze, tj. dające najmniejszy błąd klasyfikacji okazały się metody oparte na wewnętrznej zmienności cech w znalezionych skupieniach. Zastosowanie wag dla zmiennych w postaci ich wewnętrznej (lub zewnętrznej) zmienności cech w skupieniach (typ (3), (4) i (5)), może poprawiać wynik klasyfikacji przy silnej strukturze skupień. Spośród wszystkich formuł najlepszy okazał się typ (3), a wyniki pośrednie potwierdziły, że system wag równych (0) oraz standaryzacja

cech przez skalowanie (1) są zdecydowanie nieefektywne. Unitaryzacja zmiennych przy pomocy ich rozstępu (2) jest rozwiązaniem pośrednim, dającym mniejszy błąd klasyfikacji niż (0) czy (1), lecz większy niż formuły (3), (4) i (5). Na podstawie cytowanych wyników można by sądzić, że gdyby standaryzacja przez skalowanie była przeprowadzona w obrębie każdego ze znanych skupień, to jej efektywność wyodrębniania skupień byłaby wyższa. Problemem jest jednak to, że na tym etapie analizy skupień nie znamy dokładnej struktury skupień czyli przynależności obiektów do skupień.

Dla grupowania metodą Warda i metodą *k*-średnich efekty różnych formuł standaryzacyjnych badał również Steinley (2004). Formułę standaryzacyjną będącą kompromisem pomiędzy wymogiem jak najlepszego zachowania struktury skupień a nadawaniem większych wag zmiennym z większą zmiennością, zaproponowali Steinley i Brusco (por. wzór (3.37)). Formuła ta, zdaniem autorów, poprawia efektywność metody HINoV, ale ten wniosek można kwestionować (por. rozdz. 3 i 5).

Wyczerpujący przegląd formuł normalizacyjnych dla zmiennych ciągłych można znaleźć w pracy Pawełek (2008).

Kierując się cytowanymi wynikami badań, spośród zbadanych popularnych formuł standaryzacyjnych postaci

$$x_j^{'} = \frac{x_j - a}{b} \quad \text{dla} \quad j = 1, \ldots, V \tag{1.1}$$

do standaryzowania zmiennych mierzonych na skali interwałowej oraz ilorazowej wybrano formułę w postaci unitaryzacji zerowanej przy zastosowaniu rozstępu, czyli

$$x_j^{'} = \frac{x_j - \min x_j}{r_j} \quad \text{dla} \quad j = 1, \ldots, V \tag{1.2}$$

Wybór określonej miary odległości zdeterminowany jest przez skalę pomiarową, ale zależy również od sposobu standaryzacji (lub normalizacji) zmiennych. Miarą odległości powszechnie wykorzystywaną w analizie skupień w przypadku zmiennych mierzonych na silnych skalach (interwałowej i ilorazowej), ze względu na najlepiej zbadane jej własności i prostą interpretację geometryczną, jest odległość euklidesowa lub kwadrat tej odległości. Taką miarę (por. tab. 1.1) będziemy stosować dla zmiennych mierzonych na skali ilorazowej i interwałowej, w zbiorach danych występujących w eksperymentach symulacyjnych.

W przypadku zmiennych mierzonych na skali porządkowej rozsądnym wyborem, ze względu na posiadane własności jest odległość *GDM* (*Walesiak*, 2002). Odległość *GDM* dana jest wzorem

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{V} a_{ikj} b_{kij} + \sum_{j=1}^{V} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj} b_{klj}}{2\sqrt{\sum_{j=1}^{V} \sum_{l=1}^{n} a_{ilj}^2 \cdot \sum_{j=1}^{V} \sum_{l=1}^{n} b_{klj}^2}} \quad , \qquad (1.3)$$

gdzie: d_{ik} jest odległością pomiędzy obserwacjami o numerach *i, k*; symbole a_{ikj}, b_{kij}, a_{ilj}, b_{klj} są obliczane w zależności od skali pomiarowej. Dla skali ilorazowej i interwałowej stosuje się wzory

$$\begin{aligned} a_{ipj} &= x_{ij} - x_{pj} \quad \text{dla} \quad p = k, l \\ b_{krj} &= x_{kj} - x_{rj} \quad \text{dla} \quad r = i, l \end{aligned} \qquad (1.4)$$

Dla skali porządkowej stosuje się wzory

$$a_{ipj} = \begin{cases} 1 & \text{dla} \quad x_{ij} > x_{pj} \\ 0 & \text{dla} \quad x_{ij} = x_{pj} \\ -1 & \text{dla} \quad x_{ij} < x_{pj} \end{cases} \qquad (1.5)$$

dla *p=k, l,* oraz

$$b_{krj} = \begin{cases} 1 & \text{dla} \quad x_{kj} > x_{rj} \\ 0 & \text{dla} \quad x_{kj} = x_{rj} \\ -1 & \text{dla} \quad x_{kj} < x_{rj} \end{cases} \qquad (1.6)$$

dla *r=i, l.* Dla skali nominalnej stosuje się wzory

$$\sum_{j=1}^{V} \sum_{l=1}^{n} a_{ilj}^2 = \sum_{j=1}^{V} \sum_{l=1}^{n} b_{klj}^2 = V(n-1) \qquad (1.7)$$

$$a_{ikj} b_{kij} = \begin{cases} 1 & \text{dla} \quad x_{ij} = x_{kj} \\ -1 & \text{dla} \quad x_{ij} \neq x_{kj} \end{cases} \qquad (1.8)$$

oraz

$$a_{ilj} b_{klj} = \begin{cases} 1 & \text{dla} \quad x_{ij} = x_{kj} \wedge \left(x_{ij}, x_{kj} = x_{lj} \vee x_{ij}, x_{kj} \neq x_{lj} \right) \\ -1 & \text{dla} \quad x_{ij} \neq x_{kj} \wedge \begin{pmatrix} (x_{ij} \neq x_{lj}\ ,\ x_{kj} = x_{lj}) \vee \\ (x_{ij} = x_{lj}\ ,\ x_{kj} \neq x_{lj}) \vee (x_{ij}, x_{kj} \neq x_{lj}) \end{pmatrix} \end{cases}$$

(1.9)

dla $l \neq i,k$. Odległość dana wzorem (1.3) będzie stosowana zarówno w procesie grupowania obserwacji jak i przy obliczaniu wartości współczynników

korelacji. Miara *GDM* spełnia warunek symetryczności, zwrotności, nieujemności, lecz nie zawsze spełnia warunek „nierówności trójkąta". Odległość *GDM* nie zmienia swojej wartości w wyniku transformacji wartości zmiennych za pomocą dozwolonego w danej skali przekształcenia oraz jest unormowana na przedziale [0;1].

W przypadku zmiennych mierzonych na skali nominalnej (w szczególności binarnej), na ogół, stosowana jest miara Sokala-Michenera (por. tab. 1.1). Ta odległość będzie używana w niniejszej monografii.

Do pomiaru odległości pomiędzy obiektami opisanymi zmiennymi mierzonymi na różnych skalach pomiarowych zastosowana zostanie formuła (por. *Gatnar, Walesiak*, 2004, *Walesiak,* 2011)

$$d = \frac{w_1 d_1 + w_2 d_2 + w_3 d_3 + w_4 d_4}{w_1 + w_2 + w_3 + w_4} \quad , \qquad (1.10)$$

gdzie: $w_1, w_2, w_3, w_4 \in \{0,1,\ldots,V\}$ są wagami przypisanymi odległościom mierzonym na podzbiorze zmiennych ze skalą nominalną (d_1), podzbiorze zmiennych ze skalą porządkową (d_2), podzbiorze zmiennych ze skalą interwałową (d_3) oraz podzbiorze zmiennych ze skalą ilorazową (d_4). Wagi te są określone przez liczbę zmiennych odpowiadającą danej skali, wobec czego spełniają warunek $w_1 + w_2 + w_3 + w_4 = V$. Odległości d_1, d_2, d_3, d_4 obliczane będą według formuł przyjętych dla pojedynczych zmiennych.

Tabela 1.1. Miary odległości stosowane w pracy, *m* oznacza liczbę wszystkich zmiennych danego rodzaju skali w określonym podzbiorze zmiennych.

Typ skali pomiarowej	Nazwa odległości	Formuła odległości
Binarna, nominalna	Sokala-Michenera	$\frac{m_r}{m}$, m_r – liczba współrzędnych, na których oba obiekty różnią się, *m* – liczba wszystkich współrzędnych
Porządkowa	GDM	formuła (1.3) (przy *V*=*m*) z podstawieniami (1.5) i (1.6)
Ilorazowa	euklidesowa	$\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(x_i - y_i\right)^2}$, *m* – liczba wszystkich współrzędnych
Ilorazowa	GDM	formuła (1.3) (przy *V*=*m*) z podstawieniem (1.4)

Ź r ó d ł o: opracowanie własne na podstawie *Gatnar, Walesiak* (2009).

Formuła (1.10) budzi najmniej zastrzeżeń (por. *Jajuga,* 1989, *Walesiak,* 1993) spośród wszystkich prób ujednolicenia miar odległości różnych skal pomiarowych. Miary odległości pomiędzy parami obiektów, jakie zastosowano do różnych skal pomiarowych, przedstawione są w tabeli 1.1.

W literaturze przedmiotu proponowane są trzy podejścia w zakresie wyboru metod grupowania obiektów. W pierwszym wybór oparty jest na analizie własności poszczególnych metod przy wykorzystaniu informacji niezależnych od badanego zbioru danych, w drugim przy wykorzystaniu macierzy danych. Trzecie podejście zwane jest strategią grupowania, polega na syntetyzacji wyników otrzymanych przy pomocy różnych metod.

Sposób pierwszy opiera się na porównaniu wyników grupowania poszczególnych metod ze znaną przynależnością obiektów do skupień. Takie podejście wykorzystuje się w badaniu własności algorytmów grupowania w eksperymentach symulacyjnych. W przypadku empirycznych zbiorów danych liczba skupień i przynależność obiektów do skupień nie jest znana, ale znając własności algorytmów grupowania można dokonać wstępnego wyboru algorytmu. Takie podejście, nie daje całkowitej gwarancji, że wybrana metoda będzie efektywna dla konkretnego, badanego zbioru danych. Dlatego też, niezbędny jest nadzór badacza oraz odpowiednie zaprojektowanie kilkuwariantowej analizy w celu wybrania najlepszego rozwiązania.

W podejściu drugim wykorzystywana jest macierz danych. To podejście polega na formalnej ocenie cech algorytmów, wśród których wymienia się m.in. własności (*Pociecha,* 1982): 1) najlepszego obrazu – wynik grupowania nie zależy od kolejności rozważania obiektów, 2) wypukłości zbioru grupowanych obiektów, 3) połączenia obiektów na wykresie drzewa (połączenia obiektów nie przecinają się), 4) poprawnej struktury grup – gdy w wyróżnionych grupach wszystkie odległości wewnętrzne są mniejsze niż wszystkie odległości zewnętrzne (struktura grupowa) lub gdy można ustalić kolejność podobieństwa między obiektami (struktura hierarchiczna), 5) poprawnej struktury połączeń drzewa (dla procedur hierarchicznych), jeżeli rezultaty grupowania dają się przedstawić w postaci drzewa połączeń zgodnego z kolejnością podobieństwa obiektów zbioru, 6) powtarzania obiektów – dodanie kilku obiektów identycznych do już występujących w zbiorze danych, nie zmienia granicy wyróżnionych klas, 7) powtarzania grup – gdy dodanie obiektów identycznych (z już występującymi) w ramach jednego skupienia nie zmienia wyniku klasyfikacji, tzn. obiekty zostaną przypisane do tego samego skupienia, 8) opuszczania grup – po usunięciu w całości wybranego skupienia, nie zmieni się wynik grupowania, tzn. wszystkie pozostałe obiekty ponownie zostaną przydzielone do tych samych skupień, 9) monotoniczności – monotoniczna transformacja macierzy odległości nie zmienia wyników grupowania.

Tabela 1.2. Wybrane własności niektórych metod aglomeracyjnych grupowania obiektów.

Metoda aglomeracyjna	Własność:					
	wypukłości	poprawnej struktury grup	poprawnej struktury połączeń drzewa	powtarzania obiektów	opuszczania grup	monotoniczności
Pojedynczego połączenia	nie	tak	tak	tak	tak	tak
Pełnego połączenia	nie	tak	tak	tak	tak	tak
Średniej klasowej	nie	tak	tak	nie	tak	nie
Warda	tak	nie	tak	nie	tak	nie
Środka ciężkości	nie	nie	nie	nie	tak	nie

Źródło: Gordon (1999), Everitt i in. (2001), Mikulec (2010), Pociecha (1982).

Wykorzystanie własności formalnych poszczególnych procedur do wyboru najlepszej wymaga jednak ich adaptacji do badanego zbioru danych. Nie wszystkie własności podane w tabeli 1.2 są jednakowo ważne i nie wszystkimi musi cechować się stosowana metoda aglomeracyjna. Kierując się cytowanymi wynikami badawczymi, w niniejszej monografii, spośród metod aglomeracyjnych stosowane będą metody: Warda, pełnego połączenia i średniej klasowej.

Z powodu ograniczonych możliwości stosowania podejścia drugiego (opartego na macierzy danych) oraz nie dającego całkowitej gwarancji wyboru najlepszej metody grupowania podejścia pierwszego (niezależnego od danych) proponuje się trzeci sposób wyboru metod analizy skupień – strategię grupowania (por. *consensus trees, Gordon,* 1999). Ten pomysł polega na zastosowaniu różnych metod grupowania obiektów a następnie porównania uzyskanych wyników w celu wyboru rezultatu najlepszego bądź uogólnienia uzyskanych wyników. To podejście różni się nieco od dwóch poprzednich : po pierwsze, wykorzystuje się zbiór danych; po drugie, przeprowadza się kolejne wariantów grupowania dla różnej liczby skupień; po trzecie, nacisk kładzie się na ocenę uzyskanego wyniku. Podejście takie stało się popularne dzięki rozwojowi techniki komputerowej oraz dostępności oprogramowania. Mając na uwadze to wielowariantowe podejście do badanego zbioru danych analiza

skupień jest również traktowana jako najbardziej rozwinięta forma analizy kombinatorycznej (*combinatorial data analysis)*.

Metody grupowania obiektów można podzielić na kilka rodzajów (por. *Gatnar i Walesiak ,* 2004), najpopularniejsze z nich to: metody partycjonujące (podziałowe) wśród których można wyróżnić metody obszarowo-gęstościowe oraz metody optymalizujące wstępny podział zbioru danych, metody aglomeracyjne (bądź deglomeracyjne) oraz metody wizualizacji danych.

Obszerne opracowanie wyników symulacji *Monte Carlo* dla procedur aglomeracyjnych – dokonanych przez kilkunastu różnych autorów posługujących się różnej wielkości zbiorami danych pod względem liczby analizowanych cech i obiektów – w zakresie ich zdolności do wykrywania struktury skupień wygenerowanych zbiorów danych, przy występowaniu różnych czynników mogących zakłócać tę strukturę, zawiera praca Milligana i Coopera (1987). Przegląd metod wskazuje, że spośród metod aglomeracyjnych najwyżej ocenione zostały procedury pełnego wiązania, średniej grupowej, oraz Warda, przy czym na pierwszym miejscu zdecydowanie najczęściej wymieniana była metoda Warda.

Dla metod podziałowych oprócz wyboru algorytmu bardzo istotny jest dobór punktów startowych, a więc określenie k obiektów w zbiorze danych inicjujących algorytm, których wybór, na ogół, ma zasadnicze znaczenie dla uzyskanego wyniku. W przypadku metody *k*-średnich najnowsze badania (*Steinley i Brusco,* 2007) wskazują jako najlepsze rozwiązania metodę aglomeracyjną Warda, którą można zastosować do wybrania punktów startowych oraz metodę wielokrotnego losowego doboru punktów startowych i przyjęciu wyniku minimalizującego sumę kwadratów odchyleń wartości zmiennych od centrów skupień (por. wzór (1.15). Ciekawym rezultatem jest to, że lepsze wyniki uzyskuje się dzięki połączeniu metod różnych rodzajów.

Należy jednak zauważyć, że wszystkie cytowane badania przeprowadzone były przy założeniu znanej (i poprawnej) liczby skupień. Wyniki tracą więc nieco na wartości w przypadku grupowania obiektów empirycznych zbiorów danych, gdy poprawnej liczby skupień nie znamy.

Etap ustalający liczbę skupień w zbiorze danych znajduje się przed etapem grupowania obiektów ale do jego przeprowadzenia na ogół konieczne jest uzyskanie podziału zbioru na wszystkie początkowe (tj. do pewnej liczby np. 20 skupień) liczby skupień, gdyż większość znanych indeksów liczby skupień ma charakter optymalizacyjny tzn. wskazuje na optymalną liczbę skupień dla danej metody grupowania. Wśród najczęściej stosowanych wymienić należy indeksy; Bakera-Huberta, Calińskiego–Harabasza, Dunna, Daviesa-Bouldina, Hartigana, Huberta-Levine'a, Krzanowskiego-Lai, indeks sylwetkowy indeks gap. Osobną grupę tworzą indeksy opracowane tylko pod kątem metod aglomeracyjnych np.

indeks Mojeny (1977). Dla metod aglomeracyjnych, Sokołowski (1992) wyróżnia aż pięć różnych grup indeksów liczby skupień.

Efektywność wymienionych i innych indeksów badana była przez wielu autorów m.in. *Milligan i Cooper* (1985), *Migdał-Najman i Najman* (2005), *Korzeniewski* (2005) Wybór właściwych indeksów służących do oceny liczby skupień nie jest łatwy. Spośród wymienionych najlepszą opinią cieszą się indeks gap, Daviesa-Bouldina, Calińskiego–Harabasza. Słabsze od nich okazały się indeksy Dunna i Hartigana. Jak stwierdzają sami wynalazcy indeksów, niektóre z indeksów mogą okazać się nieefektywne, w przypadku niektórych zbiorów danych. Dlatego też, nie należy wyników badań traktować z całkowitą pewnością, a jednym z często stosowanych w praktyce rozwiązań jest wykorzystanie wskazań kilku indeksów z następną syntezą ich wskazań (*Gordon,* 1999). Odnotujmy wzory wykorzystywane przy obliczaniu indeksu sylwetkowego, gdyż zostaną one później użyte przy walidacji efektów analizy skupień. Wartość indeksu sylwetkowego *i*-ego obiektu dana jest wzorem

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad , \qquad (1.11)$$

gdzie wielkość $a(i) = 1/(n_k - 1) \sum_{j \in C_k \; j \neq i} d(i, j)$ to średnia odległość obiektu i od pozostałych obiektów należących do skupienia C_k (które zawiera obiekt *i*); zaś wielkość $b(i) = \min_{C_r \neq C_k} d(i, C_r)$, gdzie $d(i, C_r) = (1/n_r) \sum_{q \in C_r} d(i, q)$, można określić jako minimalną średnią odległość obiektu $i \in C_k$ od pozostałych skupień.

W literaturze etap oceny wyników grupowania jest określany jako walidacja *(validation)*, przy czym stosowane są w nim różne podejścia do analizy wyników, opierające się na: 1) testowaniu losowości obiektów (czyli braku struktury klas); 2) testowaniu kompletności struktury grup; 3) ocenie poszczególnych skupień; 4) ocenie całego wyniku grupowania; 5) ocenie struktury klasyfikacji hierarchicznej (*Gordon,* 1999). Na przykład w zakresie oceny poprawności poszczególnych skupień w ujęciu stochastycznym może być wykorzystana wartość statystyki *U* Manna-Withneya (por. *Domański,* 1979) i symulacje *Monte Carlo*. Z kolei w testowaniu uzyskanego wyniku grupowania z punktu widzenia braku struktury klas, w konstrukcji hipotezy zerowej wykorzystywane są trzy modele (por. np. *Gordon,* 1999): rozkład Poissona – w przypadku analizy czy obiekty reprezentowane przez punkty w przestrzeni *V*-wymiarowej są ułożone równomiernie w pewnym jej obszarze; rozkład jednomodalny – zakładający, że badane obiekty pochodzą właśnie z takiego *V*-wymiarowego rozkładu, tzn. tworzą jedno skupienie oraz analiza losowości

zbioru, np. macierzy odległości – dla oceny czy odległości pomiędzy obiektami są losowe, tzn. czy elementy dolnego trójkąta macierzy odległości są uporządkowane losowo. Warto zauważyć, że do oceny braku struktury klas można wykorzystać indeks przerwy (*gap index, Tibshirani i inni,* 2001), który został skonstruowany z myślą o ocenie liczby skupień w zbiorze danych ale ma tę własność, że można go stosować gdy ta liczba jest równa zeru.

W praktyce, najczęściej oceny wyniku grupowania dokonuje się za pomocą odpowiednio dobranych miar jakości klasyfikacji oraz przy pomocy replikacji klasyfikacji.

Replikacja, czyli wielokrotne powtórzenie klasyfikacji może być traktowana jako sprawdzanie krzyżowe wyniku *(cross-validation)*, dotyczy bowiem badania jak daleko identyfikacja skupień na podstawie dwóch podprób wylosowanych z analizowanego zbioru danych odpowiada ostatecznemu wynikowi grupowania, tj. przynależności obiektów do skupień otrzymanej na podstawie całego zbioru danych. W literaturze przedmiotu można znaleźć kilka propozycji miar oceny zgodności wyników grupowania (*Denoeud* i inni, 2005): indeks oparty na CER, indeks Randa, skorygowany indeks Randa, indeks Jaccarda, Wallace'a, Lermana, Fowlkesa i Mallowsa (dla metod hierarchicznych) czy wskaźnik Nowaka (1985).

Najprostszym indeksem zgodności dwóch podziałów jest chyba indeks oparty na CER (*classification error rate*) czyli odsetku błędnych klasyfikacji. Formuła tego indeksu to suma liczby par obiektów przypisanych do tego samego skupienia lub do różnych skupień w obu podziałach (czyli $t_1 + t_2$ por. (1.15)) odniesiona do liczby wszystkich par. Wartość tego indeksu jest równa wartości indeksu Randa (nieskorygowanego). Wyniki badań empirycznych dla kolejnych pięciu wymienionych indeksów w zakresie ich poprawności (dokładności) dokonane przez Denoeuda i in. wskazały na podobne zachowanie się indeksu Jaccarda, skorygowanej miary Randa, Wallace'a i Lermana. Natomiast przy szczegółowej interpretacji ich wyników spośród wymienionych miar najbardziej prawidłowymi, stabilnymi okazały się miary Jaccarda i Wallace'a, a trzecią w kolejności skorygowana miara Randa. Skorygowany indeks Randa ze względu na swoją uniwersalność oraz powszechność jego stosowania a także to, iż jest on podstawą działania innych metod, np. metody *HINoV,* będzie wykorzystywany w niniejszej pracy.

Skorygowany indeks Randa (por. *Hubert i Arabie*, 1985), dla dwóch różnych podziałów P_1, P_2 zbioru danych ma postać:

$$RI(P_1, \mathrm{P}_2) = \frac{\binom{n}{2}(t_1 + t_2) - \left[(t_1 + t_3)(t_1 + t_4) + (t_3 + t_2)(t_4 + t_2)\right]}{\binom{n}{2}^2 - \left[(t_1 + t_3)(t_1 + t_4) + (t_3 + t_2)(t_4 + t_2)\right]} \tag{1.12}$$

gdzie: t_1 - liczba par obserwacji należących do tego samego skupienia dla obu podziałów; t_2 - liczba par obserwacji należących do innych skupień dla obu podziałów; t_3 - liczba par obserwacji należących do tego samego skupienia dla podziału P_1 i do różnych skupień dla podziału P_2; t_4 - liczba symetryczna do t_3. Wskaźnik ten jest ograniczony z góry przez liczbę 1. Im bliższy 1 tym dwa porównywane podziały bardziej podobne.

Replikację opartą na indeksie Randa można prowadzać według następującego schematu (*Milligan, Cooper*, 1987):

1) Dzielimy losowo zbiór n obiektów opisanych zbiorem V cech na dwie (równoliczne) podpróby P_1 i P_2.

2) Stosujemy wybraną metodę grupowania dla podpróby P_1 by podzielić ją na znaną liczbę k skupień, a następnie wyznaczyć środki ciężkości tych skupień.

3) Wyznaczamy odległość pomiędzy każdym obiektem należącym do podpróby P_2, a centroidami skupień z podziału podpróby P_1.

4) Przypisujemy każdy obiekt z podpróby P_2 do najbliższej niemu centroidy.

5) Stosujemy tę samą metodę klasyfikacji dla podpróby P_1.

6) Porównujemy zgodność uzyskanych wyników podziału dla obydwu prób P_1 i P_2, na podstawie skorygowanego indeksu Randa.

7) Kroki 1)–6) można wielokrotnie powtórzyć, w celu wyznaczenia średniej oceny stabilności wyniku grupowania dla różnych podziałów podprób P_1 i P_2.

Jak wskazują wyniki porównań cytowane w pracy Milligana (1996) replikacja stanowi użyteczne narzędzie potwierdzające wyniki klasyfikacji, zarówno w przypadku metod aglomeracyjnych jak i podziałowych.

Obok replikacji miarą oceny jakości uzyskanego grupowania może być miara oparta na indeksie sylwetkowym zwana całkowitym indeksem sylwetkowym *GSI (Global Silhouette Index)*. Wartości indeksu *GSI* wyznacza się na podstawie cząstkowych ocen $s(i)$ przynależności każdego obiektu do swojego skupienia

$$S(C_k) = \frac{1}{n_k} \sum_{i=1}^{n_q} \sum_{i \in C_k} s(i), \qquad (1.13)$$

$$GSI = \frac{1}{K}\sum_{k=1}^{K} S(C_k). \qquad (1.14)$$

Z konstrukcji indeksu dla pojedynczego obiektu $s(i)$ wynika, że $s(i) \in \langle -1,1 \rangle$, przy czym: jeżeli obiekt i jest niepoprawnie zaklasyfikowany to $s(i) \in \langle -1,0 \rangle$; w przypadku obiektu i tworzącego skupienie jednoelementowe $s(i) = 0$; natomiast wraz ze wzrostem stopnia przynależności obiektu i do skupienia $s(i) \in (0,1\rangle$, dlatego też ostatecznie wartość $GSI \in \langle 0,1 \rangle$.

Interpretacja GSI jest następująca (*Kaufman, Rousseeuw,* 1990): wartość $GSI \geq 0{,}71$ wskazuje na silną strukturę uzyskanych grup; $GSI \in \langle 0{,}51;0{,}70 \rangle$ można uznać za poprawną strukturę grupowania; $GSI \in \langle 0{,}26;0{,}50 \rangle$ mówi, że wyodrębniona struktura jest słaba, a wynik analizy skupień może być sztuczny (być może należy zastosować inną metodę grupowania); natomiast wynik $GSI \leq 0{,}25$ nakazuje odrzucić uzyskane rozwiązania – wskazując na brak skupień w zbiorze danych.

W metodach podziałowych, dla cech ciągłych, na ogół korzysta się z wariancji jako miary jakości grupowania. Miernik jakości otrzymanego grupowania powinien uwzględniać wewnętrzną spójność otrzymanych skupień jak również zewnętrzne ich zróżnicowanie. Jeśli miernik będzie oparty na wariancji, to całkowitą wariancję należy rozdzielić na składnik wewnątrzklasowy oraz skladnik zewnątrzklasowy. Rozdzielenie wariancji całkowitej na dwa składniki, o których mowa możemy zapisać w następujący sposób:

$$\sum_{i=1}^{n}\sum_{v=1}^{V}\left(x_{iv}-\bar{x}_v\right)^2 = \sum_{k=1}^{K}\sum_{i\in C_k}\sum_{v=1}^{V}\left(x_{iv}-\bar{x}_{kv}\right)^2 + \sum_{k=1}^{K} n_k \sum_{v=1}^{V}\left(\bar{x}_v-\bar{x}_{kv}\right)^2 \qquad (1.15)$$

gdzie $\bar{x}_{kv} = \frac{1}{n_k}\sum_{i\in C_k} x_{iv}$ oznacza średnią *k*-go skupienia na *v*-ej zmiennej, zaś $\bar{x}_v$ średnią ze wszystkich obserwacji na *v*-ej zmiennej. Pierwszy składnik poprawnej stronie wzoru (1.15) jest miarą homogeniczności skupień zaś drugi jest miarą odizolowania skupień. Analogicznie można zdefiniować macierze wariancji wewnątrzklasowej i zewnątrzklasowej, które rozdzielają na poszczególne wyrazy macierzy wariancje pomiędzy parami zmiennych. Macierz wariancji zewnątrzklasowej (*between cluster variance matrix*) dana jest wzorem

$$B = \sum_{k=1}^{K} (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T, \qquad (1.16)$$

gdzie $\bar{x}_k$ oznacza wektor pionowy reprezentujący średnią k-go skupienia zaś $\bar{x}$ jest wektorem pionowym reprezentującym średnią ogólną ze wszystkich obserwacji. Macierz wariancji wewnątrzklasowej (*within cluster variance matrix*) dana jest wzorem

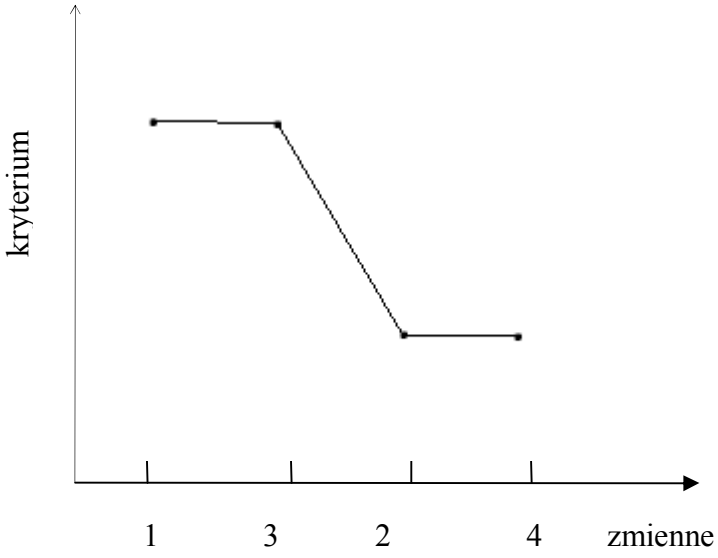
$$W = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T, \qquad (1.17)$$

gdzie x_i oznacza wektor pionowy reprezentujący *i-tą* obserwację. Na bazie tych dwóch macierzy można sformułować kryterium będące miernikiem jakości grupowania. Kryterium tym jest ślad

$$tr\left(W^{-1}B\right) \qquad (1.18)$$

Interpretacja tego śladu jest następująca: ślad (1.18) mierzy stosunek wariancji zewnątrzklasowej do wewnątrzklasowej, im wyższa jego wartość tym lepiej zdefiniowane (bardziej spójne) skupienia.

Ostatnim etapem analizy skupień jest interpretacja uzyskanych wyników, która ma trzy podstawowe cele (*Gordon*, 1999): opisanie każdego z wyodrębnionych skupień, zidentyfikowanie cech różniących poszczególne skupienia oraz umożliwienie przypisywania nowych obiektów do istniejących skupień na podstawie ich charakterystyk. Na tym etapie należy (*Gatnar, Walesiak*, 2009) wyznaczyć statystyki opisowe, tj. (o ile to możliwe) wartości średnie i odchylenia standardowe zmiennych pierwotnych w poszczególnych skupieniach celem wskazania cech charakterystycznych poszczególnych skupień. Ponadto, można wykorzystać metody graficznej wizualizacji skupień – np. dendrogramy (dla metod aglomeracyjnych) lub wykresy *silhouette plot*, obrazujące wartości indeksu sylwetkowego dla wszystkich obiektów i skupień.



Rys. 1.3. Przykład zastosowania zjawiska łokcia przy wybierania zmiennych.

Źródło: opracowanie własne.

Bardzo często spotykanym w analizie skupień kryterium jest stosowanie zjawiska tzw. „łokcia". W przypadku wybierania zmiennych idea tego zjawiska polega na tym, że mając na osi odciętych numery zmiennych a na osi rzędnych wartości ustalonego kryterium, odrzucamy te zmienne, które znajdują się „poniżej" łokcia. Na przykład, opierając się na rysunku 1.3 odrzucilibyśmy zmienne o numerach 2 i 4. Zjawiskiem tym można posługiwać się tylko w przypadku pojedynczych zbiorów danych, posługiwanie się polega na ocenie wizualnej, a więc nie jest to konkretny miernik liczbowy. Technika zjawiska łokcia pojawia się w wielu metodach analizy skupień, w odniesieniu do różnych wielkości, nie tylko przy selekcji zmiennych.

Przy wybieraniu zmiennych istotnych dla struktury skupień niezbędne jest zbadanie, o ile to możliwe, korelacji pomiędzy zmiennymi. Gdy zmienne mierzone są na skali silniejszej od nominalnej, to do pomiaru siły korelacji możemy zastosować współczynnik tau Kendalla (por. *Kendall, Buckland*, 1986, lub *Gatnar, Walesiak*, 2004) postaci

$$\Gamma_{uv} = \frac{\sum_{i=2}^{n}\sum_{k=1}^{i-1} a_{iku} b_{ikv}}{\left(\sum_{i=2}^{n}\sum_{k=1}^{i-1} a_{iku}^2 \sum_{i=2}^{n}\sum_{k=1}^{i-1} b_{ikv}^2\right)^{1/2}} \tag{1.19}$$

gdzie i, k - numer obserwacji, u, v - numer zmiennej. Ponadto, dla obserwacji mierzonych na skali ilorazowej lub przedziałowej stosujemy we wzorze (1.19) podstawienia

$$\begin{aligned} a_{iku} &= x_{iu} - x_{ku} \\ b_{ikv} &= x_{iu} - x_{kv} \end{aligned} \tag{1.20}$$

i otrzymamy wzór postaci

$$KL(u,v) = \frac{\frac{1}{n}\sum_{i=1}^{n} x_{iu} x_{iv} - \bar{x}_u \bar{x}_v}{s_u s_v} \tag{1.21}$$

gdzie $\bar{x}_u$, $\bar{x}_v$, s_u, s_v - średnie arytmetyczne i odchylenia standardowe mierzone na odpowiednich zmiennych. Dla obserwacji mierzonych na skali porządkowej stosujemy we wzorze (1.19) podstawienie (por. *Domański,* 1979)

$$a_{iku} = \begin{cases} 1 & \text{dla} \quad x_{iu} > x_{ku} \\ 0 & \text{dla} \quad x_{iu} = x_{ku} \\ -1 & \text{dla} \quad x_{iu} < x_{ku} \end{cases}, \qquad b_{ikv} = \begin{cases} 1 & \text{dla} \quad x_{iv} > x_{kv} \\ 0 & \text{dla} \quad x_{iv} = x_{kv} \\ -1 & \text{dla} \quad x_{iv} < x_{kv} \end{cases} \tag{1.22}$$

Współczynnik tau Kendalla, na ogół, przybiera wartości z przedziału $[-1,1]$.

Osobnego omówienia wymagałoby podejście modelowe do analizy skupień. Krótką charakterystykę tego działu analizy skupień zamieszczono w § 2.1.

Wobec braku możliwości szerokiego stosowania metod wnioskowania statystycznego w analizie skupień podejmowane są próby „pogodzenia" wnioskowania z metodami heurystycznymi. Taka idea przyświeca działowi analizy skupień wykorzystującemu tzw. testy empiryczne. Takie testy konstruuje się (*Sokołowski*, 1992) zakładając model populacji (na ogół mieszanki rozkładów normalnych), następnie generuje się wstępne oceny wartości krytycznych testu (hipoteza zerowa stwierdza, np. jednorodność populacji) a następnie poszukuje się funkcji analitycznej będącej dobrą aproksymacją ocen wstępnych.

1.3. Uwarunkowania procesu wybierania zmiennych w analizie skupień

Ogólnie przyjęty w literaturze pogląd (por. *Guyon i inni,* 2003, *Walesiak,* 2005) zakłada, że każde podejście do wybierania zmiennych składa się z dwóch faz. W pierwszej fazie badacz sam (lub we współpracy z ekspertami) konstruuje wstępną listę zmiennych wedle posiadanej przez siebie znajomości przedmiotu. Druga faza może mieć różny charakter. Może polegać na subiektywnej redukcji wstępnej listy zmiennych na podstawie własnej (lub ekspertów) znajomości przedmiotu badania. Druga faza może mieć również charakter bardziej formalny. Na przykład, najpierw usuwamy przy pomocy jakiegoś algorytmu zmienne o małej zmienności, następnie usuwamy zmienne powielające informacje tzn. silnie skorelowane z innymi, które zostawiamy. Ujmując rzecz ogólnie, przez metody należące do drugiej fazy rozumiemy metody statystyczne, które w możliwie największym stopniu same ustalają zbiór zmiennych tworzących strukturę (struktury) skupień. Niniejsza rozprawa poświęcona jest metodom, które należą do tej właśnie fazy.

Zagadnienie wybierania zmiennych lub inaczej redukcji wymiarowości spotykane jest w wielu dziedzinach statystyki. Korzyści jakie ze sobą niesie ewentualna redukcja wymiarowości są następujące: ułatwienie wizualizacji i zrozumienia danych, złagodzenie skutków wysokiej wymiarowości (*dimensionality curse*), zredukowanie wymagań koniecznych do pomiaru i przechowywania danych, zredukowanie czasu korzystania z danych. W analizie skupień najważniejsze są dwie pierwsze korzyści, jako że, wybieramy zmienne istotne dla struktury skupień tj. zmienne ją tworzące. Należy uwzględniać tylko te zmienne, które mają zdolność dyskryminacji poszczególnych obserwacji ze względu na przynależność do któregoś ze skupień z ich struktury istniejącej w zbiorze.

W literaturze rozróżnia się trzy grupy metod wyboru zmiennych, które mogą być zastosowane w analizie skupień:

1. Nadawanie pojedynczym zmiennym wag, które mają odzwierciedlać ich istotność dla struktury skupień.
2. Wybieranie zmiennych istotnych dla struktury skupień, które jest szczególnym przypadkiem ważenia zmiennych.
3. Transformację zbioru zmiennych oryginalnych do zbioru innych zmiennych.

W polskiej literaturze przegląd metod doboru zmiennych można znaleźć w pracy Grabińskiego (*Grabiński,* 1992). W pracy tej proponowana jest metoda ważenia zmiennych polegająca na nadawaniu wyższych wag zmiennym silniej

skorelowanym z pozostałymi zmiennymi. Należy w tym miejscu zauważyć, że nadawanie wag zmiennym nie daje wszystkich tych samych korzyści jakie daje wybieranie zmiennych istotnych i usuwanie pozostałych. Nadając wagi nie zmniejszamy rozmiarów zbiorów danych. Ponadto, jak zauważono (*Gnadadesikan i inni,* 1995), funkcje wszystkich zmiennych wraz z ich wagami są znacznie trudniejsze w interpretacji. Ponadto, nadawanie większych wag zmiennym silnie skorelowanym z pozostałymi jest zupełnie sprzeczne z redukcją wymiarowości mającą na celu usunięcie zmiennych powielających informacje, a przez to zbędnych (*redundant features*). Z kolei, w analizie skupień, podejście polegające na usuwaniu (lub zmniejszaniu wag) zmiennych powielających informacje również, nie zawsze prowadzi do właściwych rezultatów (por. *Milligan,* 1996, *Walesiak*, 2005). Ogólnie rzecz biorąc, nie można sformułować żadnych wniosków dotyczących tego czy wysoka korelacja pomiędzy pojedynczymi zmiennymi lub zbiorami zmiennych powinna mieć określony wpływ na decyzje dotyczące wyboru zmiennych – ani w kontekście uznawania zmiennych za istotne ani za nieistotne. W pracy Walesiaka (2005) pokazany jest również przykład dwóch bardzo silnie skorelowanych zmiennych, które tworzą wyraźną strukturę skupień pomimo to, że żadna z nich pojedynczo na nic podobnego nie wskazuje. Podobnych przykładów można znaleźć wiele wśród zbiorów danych generowanych dla cech ciągłych (por. 1.4.1.), zwłaszcza w przypadkach zbiorów, dla których istnieje dodatnia korelacja wewnątrzklasowa pomiędzy zmiennymi istotnymi i rozmycie skupień (*overlap*) jest wysokie. Należy zaznaczyć, że skorelowanie zmiennych ciągłych nie ma żadnego związku z tym czy te zmienne tworzą strukturę skupień – mogą tworzyć ale nie muszą. Przez strukturę skupień rozumiemy bowiem strukturę co najmniej dwóch skupień, więc konieczne jest występowanie jakichś obszarów o większej gęstości obiektów rozdzielonych obszarami rzadszymi. Zmienne o rozkładach jednomodalnych skorelowane ze sobą są powszechnie stosowane w eksperymentach symulacyjnych w roli zmiennych zanieczyszczających strukturę skupień tworzoną przez inne zmienne i ten przypadek nie jest wymysłem mającym jedynie uatrakcyjnić dyskusję akademicką – w świecie realnym, wśród zmiennych o charakterze ekonomicznym, raczej przeważają zmienne skorelowane między sobą. Spośród zmiennych makroekonomicznych wymienić można, na przykład, stopę bezrobocia, produkt krajowy brutto, inflację. Problem skorelowania zmiennych komplikuje się jeszcze bardziej przez to, że, nawet zależność funkcyjna (silna zależność stochastyczna lub statystyczna oczywiście też) pomiędzy dwiema zmiennymi, z których jedna tworzy strukturę skupień nie pozwala twierdzić, że druga zmienna jest dla tej struktury istotna (por. *Fraiman i inni,* 2008, *Raftery i Dean,* 2006). Należy bowiem pamiętać o tym, że selekcja zmiennych istotnych dla struktury skupień nie może pomijać zagadnienia obecnego we wszystkich działach statystyki, a mianowicie, redukcji

wymiarowości przez usunięcie zmiennych zbędnych ze względu na to, że powielają informacje zawarte w innych zmiennych. Jeśli dwie zmienne tworzą strukturę skupień, to każdą z nich uznajemy za istotną dla struktury skupień jeśli usunięcie jej spowodowałoby utratę struktury skupień. Jeśli usunięcie zmiennej nie powoduje utraty struktury skupień, to zmienną taką należy raczej uznać za zbędną w ogóle, nie tylko w sensie jej nieistotności dla struktury skupień, i wyeliminować. Tak więc, problem selekcji zmiennych tworzących strukturę skupień łączy się z zagadnieniem optymalizacji tej selekcji w sensie pozostawienia jak najmniejszej liczby zmiennych tworzących strukturę skupień. W tym miejscu pojawia się jednak następny problem: co zrobić gdy usunięcie zmiennej nie powoduje utraty struktury skupień ale ją zmienia? W wielu sytuacjach praktycznych bardzo duże znaczenie ma, na przykład, liczba skupień. Jeśli wystarczy informacja o tym, że dla pewnego zbioru zmiennych istnieje wyraźna struktura dwóch skupień wśród klientów banku (nie wspominając o wyraźnej strukturze trzech skupień), to menadżer banku będzie skłonny zapomnieć o wszystkich pozostałych zmiennych charakteryzujących klientów banku nawet gdyby przy udziale tych zmiennych udało się ustalić istnienie bardziej spójnej struktury większej liczby skupień. Idea eliminowania zmiennych zbędnych, które można rozpoznać poprzez silne skorelowanie z innymi zmiennymi, jest jednak w naturalny sposób sprzeczna z wykrywaniem zmiennych istotnych dla struktury skupień na podstawie ich skorelowania ze zmiennymi, o których wiemy, że tworzą strukturę skupień. Ta sprzeczność nakazuje wspomnianą ostrożność w wybieraniu zmiennych w analizie skupień. Załóżmy, na przykład, że 10 zmiennych tworzy strukturę kilku skupień. W takim przypadku, na ogół, wiele par zmiennych wykaże (być może silne) skorelowanie. Powstaje pytanie: Czy można wyeliminować zmienne silnie skorelowane z co najmniej jedną zmienną spośród pozostałych? W kontekście analizy skupień, wydaje się, że odpowiedź na to pytanie jest raczej negatywna. Takiej zmiennej raczej nie należy eliminować ponieważ może to pociągnąć za sobą zmianę struktury skupień oraz utrudnić proces wykrywania innych zmiennych tworzących strukturę. Inną dość zawiłą kwestią jest to czy powyżej sformułowane uwagi w jednakowym stopniu odnoszą się do zmiennych mierzonych na silnych skalach jak do zmiennych mierzonych na słabych skalach. Odpowiedź na to pytanie jest raczej negatywna. Niektórzy (por. *Talavera,* 2000) odwołują się jednak do silnego związku zmiennej ze wszystkimi pozostałymi zmiennymi jako do podstawy, z której wnioskują o istotności tej zmiennej dla struktury skupień. Takie rozumowanie jest uzasadnione (do pewnego stopnia, por. 4.4.2) w przypadku słabych skal pomiarowych, na których jest, na przykład tylko kilka wariantów. Inną kwestią jeszcze bardziej komplikującą problem jest rozstrzygnięcie czy dokonując selekcji zmiennych możemy założyć, że istnieje tylko jedna struktura skupień w

analizowanym zbiorze danych. Odpowiedź na to pytanie jest zdecydowanie negatywna. Niektórzy twierdzą nawet, że przy większej liczbie cech, w problemach praktycznych, regułą jest to, że istnieje wiele struktur skupień. W najogólniejszym przypadku zakłada się, że zbiory zmiennych tworzące te struktury nie muszą być rozłączne (por. *Friedman i Meulman,* 2004). Wymienione powyżej aspekty związane z problemem selekcji zmiennych w analizie skupień dowodzą tego, że problem ten jest faktycznie bardzo złożony. Trudno nawet sformułować precyzyjne kryteria orzekające o tym, że pojedyncza zmienna (lub zbiór zmiennych) jest istotna dla jakiejś struktury skupień. W monografii niniejszej ograniczymy się do wymienienia cech, którymi powinna charakteryzować się dobra metoda selekcji zmiennych w analizie skupień.

Podejścia, których ideą jest transformacja zbioru zmiennych oryginalnych takie jak analiza czynnikowa, transformacja Karhunena-Loeve czy metoda głównych składowych mają wady (por. *Dash i Liu,* 2000*,* oraz *Duda i Hart,* 1973). Prowadzą one do wyboru zmiennych (i w konsekwencji skupień), które trudno później zinterpretować oraz pozostawiają wszystkie oryginalne zmienne ponieważ są one konieczne do określenia nowych zmiennych po transformacji. Na przykład, jak pokazał Sneath (por. *Milligan,* 1996), redukcja wymiarowości przy pomocy metody głównych składowych (która ma ze swej natury możliwości zastosowania ograniczone tylko do zmiennych mierzonych na silnych skalach) może prowadzić do utraty niektórych skupień. Milligan „naiwnymi" nazywa próby transformacji pierwotnej przestrzeni klasyfikacji metodą głównych składowych w przestrzeń o mniejszej liczbie składowych ortogonalnych. Dla celów analizy skupień nie zaleca się, jako metody doboru zmiennych, stosować analizę głównych składowych, której podstawowym celem jest transformacja, a nie redukcja zmiennych. Proponowane są natomiast metody bazujące na ocenie zdolności cech do różnicowania podzbiorów obiektów lub na bezpośrednim doborze zmiennych do analizy w oparciu o ich skorelowanie. W analizie skupień należy uwzględniać tylko zmienne istotne dla struktury skupień czyli te, które pomogą wyróżnić grupy obiektów podobnych. Badania Milligana (1980) dowodzą konieczności dokonania wyboru zmiennych, gdyż już dodanie 1 lub 2 zmiennych o charakterze zakłócającym, czy pewnej frakcji obiektów nietypowych, szybko pogarsza zdolność dyskryminacji zbioru obiektów, zarówno wśród metod hierarchicznych i podziałowych. Analizę przeprowadzono na podstawie zbiorów danych o 50 obiektach wygenerowanych z 2-8 wymiarowej przestrzeni euklidesowej. Analiza obejmowała 6 rodzajów zaburzeń: 20%, 40% obiektów zakłócających, niewielkie oraz duże zaburzenie w postaci zachodzenia skupień, zakłócenie w 1 i 2 wymiarach. Porównaniem objęto 15 powszechnie używanych algorytmów aglomeracyjnych i podziałowych z losowymi oraz wskazanymi punktami startowymi. Oceny zgodności uzyskanych wyników ze znaną strukturą skupień dokonano na

podstawie indeksu Randa oraz wewnętrznej jakości dopasowania danych wejściowych do otrzymanego podziału zbioru obiektów. Wyniki świadczące o nieprzydatności metody głównych składowych w analizie skupień uzyskali również Chang (1983) (negatywna ocena kombinacji liniowych zmiennych uzyskanych metodą głównych składowych), Arabie i Hubert (1994) i De Soete i Carroll (1994) (negatywna ocena kilu pierwszych składników jako reprezentantów struktury skupień). Bardzo problematyczne jest również wykorzystanie analizy czynnikowej do wybierania zmiennych w kontekście analizy skupień (por. *Green i Krieger,* 1995, *Schaffer i Green,* 1998).

W monografii niniejszej metody służące formalnie do ważenia zmiennych są rozważane wraz z metodami służącymi do wybierania zmiennych. Ze względów wspomnianych powyżej, w pracy poświęconej wybieraniu zmiennych w kontekście analizy skupień pominięto trzecią grupę metod opartych na transformacji zmiennych oryginalnych. Przegląd i ocenę symulacyjną ośmiu najnowszych metod ukierunkowanych na analizę skupień, metod należących do dwóch pierwszych grup metod, zawiera obszerny artykuł Steinley'a i Brusco (*Steinley i Brusco,* 2008). W cytowanej pracy przyjęto warunek graniczny dla proporcji zmiennych istotnych i nieistotnych polegający na tym, że zmienne istotne powinny stanowić co najmniej 25% liczby wszystkich zmiennych. Najmniejsze odsetki zmiennych istotnych stosowane przez badaczy jakie można znaleźć w literaturze to 20% (*Dash i Liu,* 2000) oraz około 10% w jednostkowym przykładzie (*Steinley i Brusco*, 2007). Trudno natomiast znaleźć w literaturze wskazówki co do ograniczenia górnego odsetka zmiennych istotnych. W pracy a Steinley'a i Brusco (2008), ten problem jest przemilczany. Logiczne wydaje się jednak narzucenie takiego wymogu, by dobra metoda wybierania zmiennych potrafiła również rozróżnić przypadki, w których wszystkie zmienne są istotne – nie można wyrzucić żadnej. Takie przypadki analizowali m.in. Walesiak i Dudek (2008).

Istotną kwestią są próby opisania pojedynczych zmiennych przez różnego typu wskaźniki skupialności lub zdolności dyskryminacyjnej (por. § 3.1), które mają oceniać istotność pojedynczych zmiennych dla struktury skupień w zbiorze danych. Należy zaznaczyć, że wskaźniki skupialności miałyby większy sens gdyby dały się stosować się do raczej obszerniejszych zbiorów zmiennych. Im mniej liczny zbiór zmiennych tym mniejsze szanse na to, że wskaźnik pokaże istotność tego zbioru zmiennych dla struktury skupień. Najlepszym rozwiązaniem byłoby skonstruowanie takiego wskaźnika, który przyjmowałby wartości charakteryzujące się dużym, łatwym do rozróżnienia skokiem w momencie dołączania do zbioru zmiennych istotnych, zmiennej nieistotnej. O przydatności takiej cechy świadczyć mogą przykład z pracy Walesiak (2005) oraz przykład z pracy Guyon i inni (2003). Skonstruowanie takiego wskaźnika jednak nie jest łatwe, gdyż jak zauważono (por. *Guyon i inni,* 2003), niekiedy

dodanie zmiennej nieistotnej nie musi oznaczać pogorszenie jakości wskaźnika. Z drugiej jednak strony nawet najlepsze wskaźniki skupialności, mające wspomnianą cechę, nie rozwiązują do końca problemu, gdyż wymagają analizowania wszystkich możliwych podzbiorów zmiennych. Stosowanie takich wskaźników jest niemożliwe w sensie obliczeniowym, bo już dla 20 zmiennych istnieje ponad milion różnych podzbiorów tych zmiennych.

Inną bardzo istotną kwestią jest uzależnienie metod selekcji zmiennych od znajomości liczby skupień w zbiorze danych. To jest bardzo poważne ograniczenie, które w praktyce wymusza wielokrotne powtarzanie selekcji zmiennych z następnym ocenianiem wyników przy pomocy jakiegoś kryterium i wybieraniem rozwiązania najlepszego. Oczywiście, kryterium oceniające jest, na ogół, uzależnione od przyjętej metody grupowania obiektów co jest dodatkowym czynnikiem ograniczającym ogólność wyciąganych wniosków. Należy jednak zaznaczyć, że od założenia o znajomości liczby skupień nie można uwolnić się całkowicie, gdyż przy wysokiej w stosunku do liczebności zbioru danych liczbie skupień, struktura skupień staje się niemożliwa do odróżnienia od rozkładu równomiernego. Można jednak zdecydowanie ograniczyć konsekwencje tego problemu zakładając, że liczba skupień jest niewielka np. nie przekraczająca 10 skupień. W naukach społeczno-ekonomicznych takie założenie jest jak najbardziej uzasadnione – na ogół wystarcza wiedza o tym, że w interesującej nas populacji, na przykład populacji klientów firmy, jest kilka spójnych pod względem interesujących nas cech podzbiorów. Przy założeniu niewielkiej liczby skupień można próbować konstruować metody selekcji zmiennych, które nie wymagają dokładniejszej wiedzy na ten temat. Taka próba podjęta jest w niniejszej monografii.

Jednym z możliwych podziałów istniejących metod wybierania zmiennych jest podział na dwie grupy: metod: uzależnionych od przyjętej metody grupowania obserwacji (*wrapper methods*) oraz niezależnych od żadnej metody grupowania obserwacji (*filtering methods*). Formalnie rzecz biorąc ta różnica jest dużą niedogodnością dla metod pierwszego typu, która mogłaby być zniwelowana tylko tym gdyby wybieranie zmiennych dawało ten sam wynik przy zastosowaniu różnych metod grupowania obserwacji. Tego samego wyniku, oczywiście, na ogół nie uzyskuje się, co więcej dla większości metod tego typu stosowanie różnych metod grupowania obserwacji, w ogóle nie jest możliwe. Na przykład, jeśli metodę wybierania zmiennych opracowano pod kątem grupowania hierarchicznego to nie da się zastosować metod podziałowych i odwrotnie. Zatem, z formalnego punktu widzenia, ideałem byłaby metoda niewymagająca konieczności pogrupowania obserwacji. Z drugiej jednak strony metody nie odwołujące się do grupowania obserwacji mają mniejsze możliwości wyszukiwania istniejących w zbiorze danych struktur skupień.

Analizując empiryczne zbiory danych, na ogół spotykany się z różnymi skalami pomiarowymi, wobec czego nieodzownym warunkiem dobrej metody musi być możliwość stosowania jej do takich zbiorów.

Powyższe rozważania pozwalają sformułować następujące cechy jakimi powinna charakteryzować się metoda wybierania/ważenia zmiennych w analizie skupień:

- metoda powinna raczej wybierać zmienne niż nadawać wagę każdej zmiennej (choć kryteriami ostatecznymi są precyzyjniejsze kryteria podane w dalszym ciągu paragrafu);
- metoda powinna rozstrzygnąć czy w ogóle jakieś zmienne należy uznać za nieistotne dla struktury skupień tzn. nie formułujemy warunku dla górnej granicy liczby zmiennych wybranych jako istotne dla struktury skupień;
- metoda powinna rozstrzygnąć czy w ogóle jakieś zmienne należy uznać za istotne dla struktury skupień tzn. nie formułujemy warunku dla dolnej granicy liczby zmiennych wybranych jako istotne dla struktury skupień;
- metoda powinna rozstrzygnąć czy w zbiorze wszystkich zmiennych istnieją rozłączne podzbiory zmiennych, z których każdy tworzy swoją strukturę skupień;
- metoda powinna rozstrzygnąć czy w zbiorze wszystkich zmiennych istnieją niekoniecznie rozłączne podzbiory zmiennych, z których każdy tworzy swoją strukturę skupień;
- metoda nie może opierać się tylko na mierzeniu korelacji pomiędzy zmiennymi gdyż nawet bardzo wysoka wartość współczynnika korelacji liniowej pomiędzy pojedynczymi zmiennymi nie pozwala na wyciąganie żadnych wniosków na temat ewentualnej istotności zmiennych dla struktury skupień w zbiorze danych;
- wadą jest uzależnienie metody wybierania zmiennych od metody grupowania obserwacji, gdyż specyfika metody grupowania może mieć zbyt daleko idący wpływ na rezultaty dotyczące wybierania zmiennych;
- wadą jest uzależnienie metody wybierania zmiennych od znajomości liczby skupień w zbiorze danych;
- zaletą jest jeśli metoda może działać na zasadzie oceniania podzbiorów zmiennych dowolnie wybranych i o dowolnej liczebności; metoda ograniczająca się do analizowania np. pojedynczych zmiennych ma mniejsze szanse na dokonanie poprawnego wyboru zmiennych;

- selekcja zmiennych powinna być jak najbardziej optymalna tzn. należy wybrać jak najmniejszą liczbę zmiennych istotnych dla struktury (struktur) skupień;
- zaletą jest jeśli metoda może uwzględniać różne skale pomiarowe, na których mogą być mierzone zmienne opisujące obserwacje;
- metoda nie może działać wybiórczo względem określonych rodzajów rozkładów zmiennych zanieczyszczających strukturę skupień tworzoną przez inne zmienne, tzn. powinna identyfikować poprawnie zbiór zmiennych tworzących strukturę skupień dla wszystkich możliwych rodzajów zmiennych zanieczyszczających ciągłych o rozkładach jednomodalnych i równomiernych.

Metody nie wykorzystujące żadnych założeń dotyczących liczby struktur skupień nazywać będziemy metodami **bezwzględnymi** zaś metody oparte na założeniu, że w zbiorze zmiennych istnieje tylko jeden podzbiór zmiennych tworzący strukturę skupień, będziemy nazywać metodami **porównawczymi**.

Kryteriami oceny jakości selekcji zmiennych istotnych są: **pamięć** (*recall)*, **precyzja** (*precision)* oraz **odzyskiwalność** poprawnego przypisania obserwacji do skupień, w skrócie ***RI*** (*recovery,* lub *Rand index* na ogół używany) lub **asymptotyczna odzyskiwalność** poprawnego przypisania obserwacji do skupień, w skrócie *A**RI*** (*asymptotic recovery*).

Def. 1.1 Pamięć metody *M* dla zbioru danych *A* to stosunek liczby wybranych zmiennych istotnych do liczby wszystkich zmiennych istotnych.

Def. 1.2 Precyzja metody *M* dla zbioru danych *A* to stosunek liczby wybranych zmiennych istotnych do liczby wszystkich wybranych zmiennych.

Def. 1.3 Odzyskiwalność metody *M* dla zbioru danych *A* to wartość $RI(P_1, \mathrm{P}_2)$ gdzie podział P_1 to podział uzyskany w oparciu o ustaloną metodę grupowania obserwacji i wybrany podzbiór zmiennych istotnych zaś P_2 to znany podział bezbłędny zbioru na skupienia.

Def. 1.4 Asymptotyczna odzyskiwalność metody *M* dla zbioru danych *A* to średnia arytmetyczna z dużej liczby (np. 100) wartości $RI(P_1, \mathrm{P}_2)$ gdzie podział P_1 to podział uzyskany w oparciu o ustaloną metodę grupowania obserwacji i wybrany podzbiór zmiennych istotnych zaś P_2 to znany podział bezbłędny zbioru na skupienia.

W powyższych definicjach indeks Randa może być zastąpiony inną miarą podobieństwa dwóch podziałów tego samego zbioru. Pamięć, precyzję, odzyskiwalność i asymptotyczną odzyskiwalność możemy stosować tylko wtedy gdy poprawny podział wszystkich zmiennych na istotne i nieistotne jest znany. Znany musi być również poprawny czyli bezbłędny podział zbioru na skupienia. Ujęcie jakości metody w postaci asymptotycznej odzyskiwalności jest popularnym sposobem stosowanym w przypadku zmiennych ciągłych, dla których możemy wielokrotnie zastosować grupowanie np. metodą *k*-średnich z losowym wyborem punktów startowych. W przypadku zbiorów ze zmiennymi mierzonymi na słabszych skalach pomiarowych, na ogół, stosuje się grupowanie deterministyczne (np. aglomeracyjne). Wówczas otrzymujemy jedną wartość $RI(P_1, \mathrm{P}_2)$ i ją przyjmujemy za kryterium jakości selekcji zmiennych. Wymienione kryteria oceny poprawności zbioru wybranych zmiennych są niekiedy uzupełnione czwartym kryterium, którym jest odsetek bezbłędnych grupowań (% *perfect*). To kryterium jest mniej istotne, stosowane raczej tylko w eksperymentach symulacyjnych, w których oceniana metoda stosowana była do wielu zbiorów danych. Kryterium to może być pomocne w ocenie jakości metody wybierania zmiennych, gdyż asymptotyczna odzyskiwalność (lub wartość indeksu Randa) nie obejmuje wszystkich aspektów jakości metody, bo na jej wartość wpływają ewentualne wybrane zmienne nieistotne dla struktury skupień. Bezbłędną klasyfikację obiektów można otrzymać raczej tylko przy bezbłędnym wyborze zmiennych istotnych dla struktury skupień. Należy zaznaczyć, że dobór kryteriów zależy od tego czy mamy do czynienia z jednym zbiorem danych, czy z eksperymentem na wielu zbiorach danych oraz od rodzaju zmiennych opisujących obiekty. Podstawową cechą różnych mierników jakości jest to, czy zależą one od metody grupowania obiektów czy też nie. Na przykład, w eksperymencie, w którym oceniamy metodę na zbiorach danych tylko ze zmiennymi binarnymi, z dużą (>1000) liczbą obiektów w zbiorze, nie do końca sensowne jest używanie *ARI* lub indeksu Randa. Należałoby raczej skoncentrować się na miernikach odnoszących się jedynie do wybranych zmiennych, niezależnych od jakiejkolwiek metody grupowania obiektów. W literaturze przedmiotu można spotkać też kryteria oparte na ocenie stabilności grupowania obiektów otrzymanego przy uwzględnieniu tylko zmiennych wybranych przez metodę (np. indeksy stabilności oparte na indeksie sylwetkowym). Tego typu kryteria nie będą stosowane w monografii, gdyż wprowadzają one jeszcze jeden dodatkowy czynnik od którego ocena będzie uzależniona. W celu rzetelnego ocenienia metod nie można wprowadzać zbyt wielu czynników, które będą wpływały na ocenę, gdyż łączny wpływ ich wszystkich może przesłonić ocenę jakości metody. Wydaje się, że

wprowadzenie jednego czynnika jakim jest metoda grupowania (potrzebna do obliczenia indeksu Randa) jest granicą, której nie powinno się przekraczać.

1.4. Zbiory danych generowanych

1.4.1. Zbiory danych ze zmiennymi ciągłymi

Generowanie wielowymiarowych struktur skupień jest, wbrew pozorom, trudnym zadaniem. W literaturze dominuje generowanie poszczególnych skupień z rozkładów normalnych, ale nie wiadomo na ile dobrze takie modele odzwierciedlają struktury skupień z jakimi mamy do czynienia w przypadku empirycznych zbiorów danych. Ponadto, nawet jeśli założymy, że skupienia będą generowane z rozkładów normalnych, to nadal nie wiadomo jak łączyć te skupienia tak by tworzyły strukturę, której dowolny charakter można by opisać w prosty sposób przy pomocy mierników liczbowych. Cechy struktury skupień, które należałoby wziąć pod uwagę w najogólniejszym przypadku to: liczba skupień, kształt pojedynczych skupień, zróżnicowanie tego kształtu, liczebności skupień, separowalność skupień, zróżnicowanie separowalności ze względu na skupienia, zróżnicowanie separowalności ze względu na poszczególne zmienne i, być może, jeszcze inne cechy. Algorytmy generowania struktur skupień, które zaproponowano dotychczas spełniają niewielki ułamek wymienionych cech.

Wszystkie zbiory danych ze zmiennymi ciągłymi analizowane w niniejszej monografii miały struktury skupień generowane przy pomocy algorytmu OCLUS (*Steinley i Henson,* 2005) – jednego z najnowszych ze znanych z literatury sposobów. Algorytm ten jest jednym z dwóch, przy pomocy których można generować wielokrotnie zbiory danych ze strukturą skupień kontrolując stopień rozmycia skupień (*overlap,* zachodzenie skupień na siebie, pokrywanie się skupień). Drugim algorytmem jest metoda Qiu i Joe (2006), w której autorzy mierzą stopień separowalności skupień przy pomocy zaproponowanego przez nich indeksu separowalności. W swojej pracy Steinley i Henson krytykują wcześniejsze metody generowania zbiorów danych na potrzeby analizy skupień. Przez wiele lat najczęściej stosowano metodę Milligana (1985), która polega na generowaniu wyraźnie oddzielonych skupień z uciętych wielowymiarowych rozkładów normalnych a następnie dodawaniu do tych skupień obserwacji nietypowych mających za zadanie imitowanie zachodzenia skupień na siebie. Wadą tej metody jest to, że nie znamy końcowego efektu stopnia pokrywania się skupień. Ponadto, w metodzie tej stosowany jest sposób generowania polegający na wyróżnieniu niektórych współrzędnych i narzuceniu na nie innych warunków dotyczących stopnia pokrywania się skupień niż dla pozostałych współrzędnych. Ta cecha może niekiedy być uważana za dobre naśladownictwo rzeczywistości

ale raczej nie w tym stopniu w jakim stosowana jest w tym algorytmie. Autor zakłada bowiem, że na pierwszej współrzędnej skupienia są niemal doskonale separowalne, a na pozostałych mogą zachodzić na siebie. Wada niemożności kontrolowania końcowego stopnia pokrywania się skupień dotyczy również metody Kuipera i Fishera (1975), która polega na tym by generować skupienia z wielowymiarowych rozkładów normalnych zmieniając ich średnie i macierze kowariancji. Gold i Hoffman (1976) generowali zbiór podstawowy z wielowymiarowych rozkładów normalnych, a następnie dodawali obserwacje wylosowane ze zmiennych o różnych wektorach średnich. Taka metoda również nie pozwala na określenie końcowego efektu stopnia pokrywania się skupień. Blashfield (1976) generował skupienia z rozkładów, które mogły być skorelowane ze sobą, a następnie dodawał obserwacje nietypowe z rozkładu równomiernego. Ta metoda również nie pozwala na określenie końcowego efektu stopnia pokrywania się skupień, ponieważ kilka parametrów jest wybieranych losowo. McIntyre i Bashfield (1980) zmieniali zróżnicowanie rozkładów, z których generowano skupienia ale, podobnie jak we wcześniejszych metodach, nie określili precyzyjnie stopnia pokrywania się skupień. Price (1993) kontrolował stopień pokrywania się skupień metodą prób i błędów tj. przesuwając średnie rozkładów do momentu powstania wymaganego stopnia. Ta metoda jest bardzo czasochłonna i nie wszystkie stopnie pokrywania się skupień można przy jej pomocy uzyskać. Atlas i Overall (1994) generowali skupienia kontrolując stopień pokrywania się ich ale, jak sami zauważyli, dla całego zbioru stopień ten nie był znany. Podobnie, w metodzie Wallera i innych (1999), można kontrolować stopień pokrywania się skupień ale tylko przy niskim wymiarze przestrzeni oraz wizualnie co wyklucza tę metodę jeśli chcemy generować dużą liczbę zbiorów.

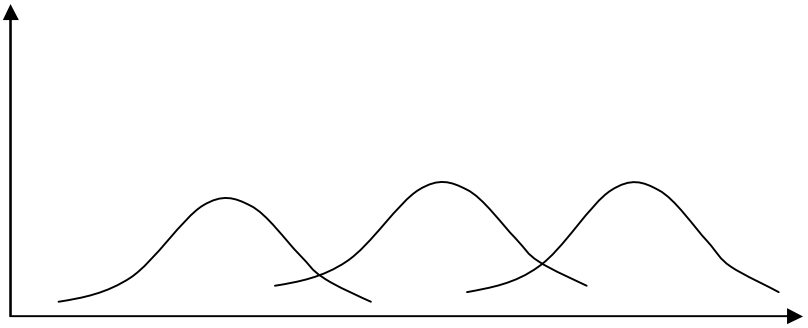
W algorytmie OCLUS separowalność skupień jest typu "łańcuchowego" (por. rys. 1.4) wobec czego na każdym wymiarze jest k-1 par skupień zachodzących na siebie w takim samym stopniu równym $(overlap)^{1/T}$ (k – liczba skupień, T – liczba wymiarów). Ogólny stopień pokrywania się skupień jest bowiem iloczynem stopni pokrywania się skupień na poszczególnych wymiarach, czyli

$$overlap = \prod_{i=1}^{T} overlap_i \qquad (1.22)$$

Z kolei, stopień pokrywania się skupień na każdym wymiarze (*marginal overlap*) jest równy średniej arytmetycznej ze wszystkich stopni pokrywania się sąsiednich skupień, czyli

$$overlap_i = \frac{1}{k-1} \sum_{j=1}^{k-1} p_{j-1,j} \qquad (1.23)$$

Przyjmujemy, że każde dwa sąsiednie skupienia zachodzą na siebie w jednakowym stopniu, czyli tzw. łańcuchowy typ pokrywania się skupień. Założenie o łańcuchowym typie zachodzenia na siebie skupień jest dość silnym założeniem, gdyż nie dopuszcza takiej sytuacji, że niektóre skupienia sąsiadujące ze sobą bardziej na siebie zachodzą, inne mniej. Jednak dopuszczenie możliwości zróżnicowanego zachodzenia skupień na siebie w połączeniu z różnymi konfiguracjami liczebności skupień i różnymi liczbami skupień rozbudowałoby eksperyment do olbrzymich rozmiarów. Poszczególne skupienia są zatem generowane na każdej współrzędnej z rozkładów normalnych o jednostkowej wariancji i wartościach średnich zdeterminowanych przez stopień zachodzenia na siebie skupień oraz przez wartość średnią rozkładu generującego pierwsze skupienie. Warunek jednakowej wariancji we wszystkich skupieniach jest w algorytmie OCLUS



Rys.1.4 Schemat konstrukcji brzegowego rozkładu prawdopodobieństwa według łańcuchowego typu zachodzenia skupień na siebie dla 3 równolicznych skupień.

Źródło: opracowanie własne.

warunkiem koniecznym do tego, by można było ocenić stopień separowalności skupień. Wartość średnia (centrum) pierwszego skupienia jest równa 0. Rozkład normalny generowany był według algorytmu Marsaglia-Bray (por. *Wieczorkowski, Zieliński*, 1997) z generatorem liczb pseudolosowych z języka Delphi4. Następnie, po wygenerowaniu wszystkich skupień (na każdej współrzędnej oddzielnie) zostały one w sposób losowy ponumerowane w sposób niezależny na każdej współrzędnej.

W tabeli 1.3 podane są odległości pomiędzy centrami kolejnych skupień w zależności od stopnia zachodzenia skupień na siebie oraz liczby zmiennych. Dla *overlap*=0 przyjęto, że centra dwóch sąsiednich skupień są odległe o 6 odchyleń standardowych. Taki sposób generowania struktury skupień powoduje to, że otrzymujemy zbiory, których struktura jest efektem „równego” udziału wszystkich zmiennych. Tę cechę można uważać za wadę, wśród zbiorów danych

ze świata realnego takiego równouprawnienia cech zapewne nie ma. Z drugiej jednak strony trudniej jest modelować zdominowanie jednych zmiennych istotnych przez inne zmienne istotne. Nie wiadomo bowiem do jakiego stopnia można pójść z dominacją. Gdybyśmy poszli za daleko, to metody wybierania zmiennych istotnych mogłyby odrzucać mające mniejsze znaczenie zmienne istotne.

Tabela 1.3 Odległości pomiędzy dwoma sąsiednimi centrami skupień, w zależności od stopnia pokrywania się skupień (*overlap*) oraz liczby *T* zmiennych opisujących obiekty.

overlap	*T*=2	*T*=4	*T*=6	*T*=25	*T*=50	*T*=125
0	6	6	6	6	6	6
0,1	2	1,16	0,82	0,22	0,11	0,04
0,2	1,52	0,86	0,60	0,16	0,07	0,03
0,3	1,2	0,66	0,46	0,12	0,05	0,015
0,4	0,96	0,51	0,35	0,10	0,04	0,008

Źródło: opracowanie własne.

Przy zastosowanym sposobie generowania struktur skupień wraz z rosnącą liczbą zmiennych maleją odległości pomiędzy sąsiednimi skupieniami. Można to łatwo zaobserwować w wierszach tabeli 1.3. Łatwo widzieć, że przy dość dużej liczbie zmiennych na pojedynczych wymiarach lub nawet na dwu czy trzyelementowych podzbiorach zmiennych, struktury skupień są nierozróżnialne. Można je rozróżnić dopiero wtedy, gdy weźmiemy pod uwagę większą liczbę zmiennych, sumując w ten sposób informacje z pojedynczych zmiennych. Metody wybierania zmiennych oparte na analizowaniu pojedynczych zmiennych, par zmiennych lub nawet trójek są przy większej liczbie zmiennych skazane na porażkę.

W celu zachowania ogólnej porównywalności przeprowadzanych eksperymentów symulacyjnych wygenerowano kilka zestawów zbiorów przy pomocy algorytmu OCLUS, podobnych do zestawu zbiorów wygenerowanych przez Steinley'a i Brusco (2008). Eksperyment Steinley'a i Brusco ma jednak wiele mankamentów, wśród których należy wymienić następujące:

- do każdej struktury skupień dołączano tylko jeden wariant kombinacji zmiennych nieistotnych co mogło mieć wpływ na otrzymane wyniki;

- nie uwzględniono struktur składających się z 2 oraz 3 skupień; takie struktury są często spotykane w praktyce, a ponadto, często są trudniejsze do wykrycia od struktur większej liczby skupień;
- wydaje się, że niepotrzebnie uwzględniano bardzo małe zbiory danych z dwiema zmiennymi istotnymi i dwiema nieistotnymi; tak małe zbiory można dokładnie zbadać przy pomocy metod wizualizacji danych i badacz z danej dziedziny sam podejmie decyzję o tym które zmienne są istotne;
- nie zastosowano żadnego kryterium, które uwzględniałoby zdolności metod do podejmowania decyzji bezwzględnych a nie tylko porównawczych (najlepsze są zdaniem autorów metody, które potrafią podzielić zbiór wszystkich zmiennych na dwa podzbiory i wybrać lepszy z nich);
- nie rozważano w żadnej formie wielokrotnych struktur danych, nawet przy założeniu rozłączności zbiorów zmiennych je tworzących;
- autorzy nie podają szczegółów sposobu generowania zbiorów danych wobec czego wyniki nie są do końca porównywalne z wynikami innych badań i eksperyment wymaga powtórzenia przynajmniej dla tych metod, które okazały się najlepsze.

W zestawie zbiorów generowanych przez nas zakres niektórych parametrów eksperymentu został poszerzony, dołączono zbiory z 2 i 3 skupieniami i nieistotne rozkłady równomierne. Wszystkie zbiory składały się z 200 elementów podzielonych na kilka skupień, które na każdej współrzędnej były generowane z rozkładów normalnych z jednostkową wariancją. Poszczególne parametry zestawu podstawowego (będziemy go nazywać PODST) wygenerowanych zbiorów danych przedstawiają się następująco.

Parametr pierwszy, liczba skupień mogła być równa 2, 3, 4, 6 lub 8.

Parametr drugi, liczebności skupień, miał trzy warianty: (a) równe liczebności wszystkich skupień; (b) 10% obserwacji i (c) 60% obserwacji w jednym skupieniu, a pozostałe skupienia równoliczne.

Parametr trzeci, liczba zmiennych istotnych : 2, 4 lub 6.

Parametr czwarty, stopień pokrywania się skupień na każdej ze zmiennych istotnych miał pięć wariantów – 0; 0,1; 0,2; 0,3; 0,4.

Parametr piąty, siła korelacji wewnątrz skupień miała dwa warianty: (a) macierz kowariancji w każdym skupieniu była macierzą jednostkową; (b) w każdym skupieniu była taka sama macierz kowariancji z jedynkami na przekątnej zaś poza przekątną, liczbą wylosowaną z przedziału [0,3; 0,8].

Parametr szósty, liczba zmiennych zakłócających była równa – 2, 4 lub 6.

Parametr siódmy, rozkład zmiennych nieistotnych miał siedem wariantów : (a) wszystkie zmienne miały niezależne rozkłady skośne jednomodalne (rozkład gamma z jednym stopniem swobody dla licznika i jednym dla mianownika); (b) wszystkie zmienne miały niezależne rozkłady normalne ze średnimi równymi 20 i jednostkową macierzą kowariancji; (c) wszystkie zmienne miały

wielowymiarowy rozkład normalny ze średnimi równymi 20 i jedynkami na przekątnej w macierzy kowariancji i 0,25 poza przekątną; (d) wszystkie zmienne miały wielowymiarowy rozkład normalny ze średnimi równymi 20 i jedynkami na przekątnej w macierzy kowariancji i 0,5 poza przekątną; (e) wszystkie zmienne miały wielowymiarowy rozkład normalny ze średnimi równymi 20 i jedynkami na przekątnej w macierzy kowariancji i 0,75 poza przekątną; (f) wszystkie zmienne miały niezależne rozkłady normalne ze średnimi równymi 20 i wariancjami losowanymi z odcinka [1; 20] ; (g) wszystkie zmienne miały niezależne rozkłady równomierne na odcinku [1;40]. Razem wszystkie kombinacje parametrów dały liczbę 9450 zbiorów.

Na potrzeby wstępnego badania niektórych metod, które ma za zadanie dać odpowiedź na pytanie czy konieczne jest przeprowadzanie pełnego badania na zestawie PODST, wygenerowany został znacznie węższy zestaw zbiorów. Zestaw ten oznaczmy przez PROB. Składa się on z 96 zbiorów, które różnią się następującymi cechami. Liczba zmiennych istotnych jest taka sama jak zakłócajacych i jest równa 4, stopień pokrywania się skupień jest równy 0, 0,1, 0,2, lub 0,3; liczba skupień jest równa 2, 4 lub 6, liczebności wszystkich skupień są jednakowe, wszystkie zmienne nieistotne mają ten sam rozkład, w 4 różnych wariantach odpowiadających wariantom (a), (b), (e), (g) z zestawu PODST; siła korelacji ma dwa warianty takie same jak w zestawie PODST. Zestaw zbiorów PROB jest znacznie uproszczony w stosunku do zestawu PODST, wobec czego, jeżeli badana metoda wypadnie słabo w eksperymencie na zestawie PROB, to nie ma potrzeby przeprowadzania pełnego badania na zestawie PODST.

Po wygenerowaniu zbiorów danych z zestawu PODST zastosowano formułę unitaryzacji zerowanej (1.2).

1.4.2. Zbiory danych ze zmiennymi porządkowymi

Sztuczne zbiory danych porządkowych spotyka się w literaturze rzadko. Powodem jest zapewne to, że dane dyskretne stwarzają bardzo dużą dowolność tworzenia struktur skupień, gdyż, formalnie rzecz biorąc każdy układ wartości zmiennej dyskretnej można uznać za strukturę skupień, jeśli nie określimy żadnych założeń dodatkowych dotyczących, na przykład, liczby skupień. Ale wygenerowanie jakiegoś zestawu zbiorów danych porządkowych, najlepiej dość ogólnej postaci, wydaje się koniecznością, gdyż dobra efektywność nowej metody selekcji zmiennych w eksperymencie na takim zbiorze jest warunkiem koniecznym, który należy sprawdzić zanim zaczniemy oceniać metodę w zastosowaniu do zbiorów empirycznych. W celu zachowania porównywalności wyników zbiory danych ze zmiennymi porządkowymi analizowane w niniejszej rozprawie miały struktury skupień generowane tak jak w eksperymencie

przeprowadzonym przez Walesiaka i Dudka (2007). Najpierw wygenerowano zmienne ciągłe zgodnie z następującymi 10 modelami.

Model 1. Liczebność zbioru n=100, liczba zmiennych V=5, liczba zmiennych zakłócających M=3, liczba zmiennych istotnych T=2, liczba skupień K=2, wektor liczebności skupień to [50, 50]. Obserwacje na zmiennych istotnych wygenerowane zostały z dwuwymiarowego rozkładu normalnego ze średnimi dla poszczególnych skupień, odpowiednio [0, 0], [1, 5] i macierzą kowariancji taką samą dla wszystkich skupień z 1 na przekątnej głównej i -0.9 poza przekątną.

Model 2. Liczebność zbioru n=120, liczba zmiennych V=7, liczba zmiennych zakłócających M=5, liczba zmiennych istotnych T=2, liczba skupień K=3, wektor liczebności skupień to [60, 30, 30]. Obserwacje na zmiennych istotnych wygenerowane zostały z dwuwymiarowego rozkładu normalnego ze średnimi dla poszczególnych skupień, odpowiednio [0; 0], [1,5; 7], [3; 14] i macierzą kowariancji taką samą dla wszystkich skupień z 1 na przekątnej głównej i -0.9 poza przekątną.

Model 3. Liczebność zbioru n=140, liczba zmiennych V=10, liczba zmiennych zakłócających M=7, liczba zmiennych istotnych T=3, liczba skupień K=3, wektor liczebności skupień to [70, 35, 35]. Obserwacje na zmiennych istotnych wygenerowane zostały z trójwymiarowego rozkładu normalnego ze średnimi dla poszczególnych skupień, odpowiednio [1,5; 6; -3], [3; 12; -6], [4,5; 18; 9] i macierzą kowariancji taką samą dla wszystkich skupień z 1 na przekątnej głównej i -0.9 poza przekątną, za wyjątkiem kowariancji $\sigma_{23} = 0.9$.

Model 4. Liczebność zbioru n=125, liczba zmiennych V=3, liczba zmiennych zakłócających M=1, liczba zmiennych istotnych T=2, liczba skupień K=5, wektor liczebności skupień to [25, 25, 25, 25, 25]. Obserwacje na zmiennych istotnych wygenerowane zostały z dwuwymiarowego rozkładu normalnego ze średnimi dla poszczególnych skupień, odpowiednio [5, 5], [-3, 3], [3, -3], [0; 0], [-5; -5] i macierzą kowariancji taką samą dla wszystkich skupień z 1 na przekątnej głównej i -0.9 poza przekątną.

Model 5. Liczebność zbioru n=150, liczba zmiennych V=5, liczba zmiennych zakłócających M=2, liczba zmiennych istotnych T=3, liczba skupień K=5, wektor liczebności skupień to [30, 30, 30, 30, 30]. Obserwacje na zmiennych istotnych wygenerowane zostały z trójwymiarowego rozkładu normalnego ze średnimi dla poszczególnych skupień, odpowiednio [5; 5; 5], [-3; 3; -3], [3; -3; 3], [0; 0; 0], [-5; -5; -5] i macierzą kowariancji taką samą dla wszystkich skupień z 1 na przekątnej głównej i 0.9 poza przekątną.

Model 6. Liczebność zbioru n=130, liczba zmiennych V=10, liczba zmiennych zakłócających M=8, liczba zmiennych istotnych T=2, liczba

skupień K=5, wektor liczebności skupień to [50, 20, 20, 20, 20]. Obserwacje na zmiennych istotnych wygenerowane zostały z dwuwymiarowego rozkładu normalnego ze średnimi dla poszczególnych skupień, odpowiednio [0; 0], [0; 10], [5; 5], [10; 0], [10; 10] i jednostkową macierzą kowariancji taką samą dla wszystkich skupień z 1 na przekątnej głównej i 0 poza przekątną.

Model 7. Liczebność zbioru n=150, liczba zmiennych V=9, liczba zmiennych zakłócających M=6, liczba zmiennych istotnych T=3, liczba skupień K=5, wektor liczebności skupień to [30, 30, 30, 30, 30]. Obserwacje na zmiennych istotnych wygenerowane zostały z trójwymiarowego rozkładu normalnego ze średnimi dla poszczególnych skupień, odpowiednio [0; 0; 0], [10; 10; 10], [-10; -10; -10], [10; -10; 10], [-10; 10; 10] i macierzą kowariancji taką samą dla wszystkich skupień z 3 na przekątnej głównej i 2 poza przekątną.

Model 8. Liczebność zbioru n=150, liczba zmiennych V=6, liczba zmiennych nieistotnych M=4, liczba zmiennych istotnych T=2, liczba skupień K=4, wektor liczebności skupień to [50, 50, 25, 25]. Obserwacje na zmiennych istotnych wygenerowane zostały z dwuwymiarowego rozkładu normalnego ze średnimi dla poszczególnych skupień, odpowiednio [-4; 5], [5; 14], [14; 5], [5; -4] i jednostkową macierzą kowariancji taką samą dla wszystkich skupień z 1 na przekątnej głównej i 0 poza przekątną.

Model 9. Liczebność zbioru n=120, liczba zmiennych V=12, liczba zmiennych zakłócających M=9, liczba zmiennych istotnych T=3, liczba skupień K=4, wektor liczebności skupień to [35, 35, 35, 35]. Obserwacje na zmiennych istotnych wygenerowane zostały z trójwymiarowego rozkładu normalnego ze średnimi dla poszczególnych skupień, odpowiednio [-4; 5; -4], [5; 14; 5], [14; 5; 14], [5; -4; 5] i jednostkową macierzą kowariancji taką samą dla wszystkich skupień z 1 na przekątnej głównej i 0 poza przekątną.

Model 10. Liczebność zbioru n=140, liczba zmiennych V=10, liczba zmiennych zakłócających M=9, liczba zmiennych istotnych T=1, liczba skupień K=4, wektor liczebności skupień to [35, 35, 35, 35]. Obserwacje na zmiennych istotnych wygenerowane zostały niezależnie z jednowymiarowych rozkładów normalnych ze średnimi dla poszczególnych skupień, odpowiednio -2, 4, 10, 16 i jednakowymi wariancjami równymi 0,5.

Wszystkie obserwacje na zmiennych zakłócających wygenerowane zostały niezależnie z rozkładu jednostajnego na odcinku [0; 2,5]. Następnie wartości wszystkich obserwacji zostały przekształcone do wartości ze zbioru {1, 2, 3, 4, 5} w zależności od tego do którego z 5 przedziałów należały. Przedziały miały rozpiętość równą 1/5 obszaru zmienności $r(v)$ każdej zmiennej, były rozłączne i pokrywały cały obszar zmienności każdej zmiennej.

1.4.3. Zbiory danych ze zmiennymi binarnymi

W celu zachowania porównywalności wyników zbiory danych ze zmiennymi binarnymi analizowane w niniejszej rozprawie miały struktury skupień generowane tak jak w eksperymencie przeprowadzonym przez Brusco (2004). W eksperymencie tym analizowane były zbiory z liczbą skupień $K = 4$, 6, 8 oraz liczbą zmiennych istotnych równą $T = 4, 6, 8$.

W tabeli 1.4 przedstawione są wartości zmiennych dla poszczególnych skupień. Wektory wartości reprezentujące poszczególne skupienia zostały tak dobrane by skupienia były jak najbardziej rozróżnialne pomiędzy sobą. Pozostałe parametry eksperymentu przestawione są w tabeli 1.5. Obserwacje generowane były w ten sposób, że losowo wybrana liczba z odcinak [0, 1] decydowała o tym, do którego skupienia ma należeć kolejny generowany obiekt zgodnie z wektorem liczebności skupień podanym w tabeli 1.5. Po ustaleniu

Tabela 1.4 Wektory określające wartości zmiennych w poszczególnych skupieniach. Wiersze reprezentują skupienia, kolumny reprezentują kolejne zmienne.

Liczba skupień	Liczba zmiennych		
	4	6	8
4	1 0 0 1	1 0 0 1 1 0	1 0 0 1 1 0 1 0
	1 1 1 0	1 1 1 0 0 0	1 1 1 0 0 0 1 0
	0 0 1 1	0 0 1 1 0 0	0 0 1 1 0 0 0 0
	0 1 0 1	0 1 0 1 0 1	0 1 0 1 0 1 1 0
6	1 0 0 1	1 0 0 0 1 1	1 0 0 0 1 1 0 1
	1 1 1 1	1 1 0 1 1 0	1 1 0 1 1 0 1 0
	1 0 1 0	1 1 1 0 0 0	1 1 1 0 0 0 0 1
	0 1 0 1	0 1 0 0 0 1	0 1 0 0 0 1 1 1
	0 0 0 1	0 1 1 1 1 0	0 1 1 1 1 0 1 1
	0 1 1 0	0 0 0 1 1 0	0 0 0 1 1 0 0 1
8	1 0 1 1	1 0 0 1 1 1	1 0 0 1 1 1 0 1
	1 0 0 0	1 0 1 0 0 0	1 0 1 0 0 0 1 1
	1 1 1 0	1 1 1 1 1 1	1 1 1 1 1 1 0 0
	1 1 0 1	1 1 0 0 0 1	1 1 0 0 0 1 0 1
	0 1 0 1	0 1 0 0 1 0	0 1 0 0 1 0 0 1
	0 1 0 0	0 1 1 0 0 1	0 1 1 0 0 1 0 1
	0 0 1 1	0 0 1 1 1 0	0 0 1 1 1 0 1 0
	0 0 0 1	0 0 1 0 0 1	0 0 1 0 0 1 0 1

Źródło: Brusco (2004)

skupienia obiekt otrzymywał takie wartości zmiennych na jakie wskazują wektory z tabeli 1.4. Stopień rozmycia skupień (odpowiednik *overlap* dla zmiennych ciągłych) należy rozumieć w ten sposób, że po wygenerowaniu wartości wszystkich obserwacji na wszystkich zmiennych istotnych wybrano losowo ze wszystkich nT zer lub jedynek wskazany procent elementów (tylko w modelu 2 i 3) i jedynki zamieniono na zera zaś zera na jedynki. Wartości wszystkich obserwacji na zmiennych nieistotnych wygenerowane zostały losowo.

Wygenerowano ponadto trzy zbiory danych binarnych z podwójną strukturą skupień. Każdy zbiór składał się z 18 zmiennych, 6 tworzących jedną strukturę skupień, 6 tworzących drugą strukturę niezależną od pierwszej i 6 zmiennych nieistotnych o rozkładach równomiernych. Obie struktury skupień składały się z 6 skupień i dane były wektorami takimi jak w tabeli 1.4 (6 zmiennych, 6 skupień) z równolicznymi skupieniami. Pierwszy zbiór miał stopień rozmycia obu skupień równy 0 drugi 3% (tak jak w *Brusco* 2004), zaś trzeci 5% ,o wiele wyższy od stopnia rozmycia przyjętego przez Brusco (2004).

Tabela 1.5 Wektory określające wartości zmiennych w poszczególnych skupieniach. Wiersze reprezentują skupienia, kolumny reprezentują kolejne zmienne.

	Model		
	1	2	3
Liczebność n zbioru	2000	3000	4000
Liczba skupień	4	6	8
Liczba T zmiennych istotnych	4	6	8
Liczebności skupień (względne w relacji do n)	równe	[0,375; 0,375; 0,125; 0,125] [0,25; 0,25; 0,25; 0,083; 0,083; 0,083] [0,1875; 0,1875; 0,1875; 0,1875; 0,0625; 0,0625; 0,0625; 0,0625;]	[0,5; 0,25; 0,15; 0,1] [0,3; 0,25; 0,2; 0,125; 0,075; 0,05] [0,25; 0,2; 0,175; 0,125; 0,075; 0,065; 0,06; 0,05;]
Stopień rozmycia skupień	0%	2%	4%
Liczba M zmiennych nieistotnych	0	4	8

Źródło: Brusco (2004)

2. Metody modelowe selekcji zmiennych w analizie skupień

2.1. Uwagi wstępne

Punktem wyjścia do większości metod modelowych jest założenie o tym, że zbiór danych traktujemy jako próbę wylosowaną z populacji, której rozkład f modelujemy przy pomocy mieszaniny rozkładów f_k gdzie k jest numerem skupienia. Skupienia są w podejściu modelowym identyfikowane z gęstościami f_k. A zatem

$$f(x) = \sum_{k=1}^{K} \alpha_k f_k(x) \tag{2.1}$$

przy czym $\alpha_k \geq 0$, $\sum_{k=1}^{K} \alpha_k = 1$. Modele mieszanek tej postaci są stosowane w statystyce od dawna, po raz pierwszy były użyte przez Karola Pearsona (1894). Gdy wprowadzimy parametry θ_k k-go skupienia, to otrzymamy zapis

$$f(x) = \sum_{k=1}^{K} \alpha_k f_k(x, \theta_k) \quad . \tag{2.2}$$

Opracowanych zostało kilka metod wykorzystujących postać (2.2) modelu populacji. Równanie (2.2) wykorzystuje się do modelowania populacji opisanych cechami ciągłymi, na ogół zakładając rozkład normalny dla poszczególnych składników sumy. Głównym celem modelowego podejścia w analizie skupień jest ustalenie liczby skupień, liczebności skupień, oszacowanie parametrów każdego skupienia i przypisanie obiektów do właściwych skupień. Kilka lat temu pojawiły się również pierwsze propozycje wykorzystania modelu tej postaci do selekcji zmiennych tworzących strukturę skupień. Funkcja wiarygodności potrzebna do estymacji wektora parametrów $\Theta = [\theta_1, \ldots, \theta_K]^T$ ma postać

$$L(\Theta | x) = \prod_{i=1}^{n} \sum_{k=1}^{K} \alpha_k f_k(x_i | \theta_k) \tag{2.3}$$

Należy znaleźć wartości wszystkich parametrów (α_k również), by funkcja (2.3) osiągnęła wartość maksymalną. Jeżeli, zgodnie z ogólnie przyjętymi oznaczeniami (por. załącznik), włączymy do wzoru (2.3) wektor z_i , to, po zlogarytmowaniu otrzymamy postać

$$L(\Theta, \alpha_k, z_{ik} | x) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\alpha_k f_k(x_i | \theta_k)) \tag{2.4}$$

Do szacowania nieznanych parametrów stosowany jest algorytm EM, który składa się z dwóch kroków. W kroku E wyznaczane są wartości estymatorów $\hat{z}_{ik}$

$$\hat{z}_{ik} = \frac{\hat{\alpha}_k f_k\left(x_i \middle| \hat{\theta}_k\right)}{\sum_{l=1}^{K} \hat{\alpha}_l f_l\left(x_i \middle| \hat{\theta}_l\right)} \tag{2.5}$$

W kroku M obliczane są wartości estymatorów $\hat{\alpha}_k$ oraz $\hat{\theta}_k$. Kroki E oraz M stosowane są na przemian do momentu, w którym wartości estymatorów nie zmieniają się (lub zmieniają się w małym stopniu).

W dalszym ciągu rozdziału scharakteryzowane zostaną trzy metody bazujące na modelowaniu populacji w postaci mieszanek rozkładów normalnych. Wśród metod nieprezentowanych w tym rozdziale specyficzną grupę tworzą metody (np. *Tadessse i inni* , 2005, *Tadessse i inni* , 2006, *Hoff*, 2006) opracowane pod kątem zbiorów danych, w których liczba zmiennych jest dużo wyższa od liczby obserwacji ($V >> n$). Metody te znajdują zastosowanie w genetyce i medycynie (analiza łańcuchów DNA). Cechą tych metod jest podejście bayesowskie zakładające, że parametr θ_k ma rozkład Dirichleta. Możliwości zastosowań tego rodzaju metod do zbiorów danych o charakterze ekonomiczno-socjologicznym są ograniczone.

2.2. Metoda wyrazistości cech

Law wraz z innymi badaczami (*Law i inni* , 2004) zaproponował wersję modelu, zakładając, że każde skupienie ma wielowymiarowy rozkład normalny z diagonalna macierzą kowariancji (co jest rodzajem warunkowej niezależności zmiennych). To założenie pozwala zapisać gęstość (2.2) w postaci

$$f(x) = \sum_{k=1}^{K} \alpha_k \prod_{v=1}^{V} f_k(x_v | \theta_{kv}) \tag{2.6}$$

Następnie, w celu uchwycenia wyrazistości cech (*feature saliency*), zbiór wszystkich zmiennych podzielono na dwa podzbiory: zmiennych istotnych i nieistotnych. Rozkład zmiennych nieistotnych nie zależy od numeru skupienia, wobec tego, jeśli oznaczymy rozkład na v-ej zmiennej (nieistotnej) przez $q(x_v | \lambda_v)$, to wzór (2.6) będziemy mogli zapisać w postaci

$$f(x)=\sum_{k=1}^{K}\alpha_k\prod_{v=1}^{V}\left[f_k\left(x_v|\theta_{kv}\right)\right]^{\phi_v}\left[q\left(x_v|\lambda_v\right)\right]^{1-\phi_v} \qquad (2.7)$$

gdzie ϕ_v będzie parametrem binarnym decydującym o tym czy zmienna jest istotna dla struktury skupień czy nie. Jeśli $\phi_v = 0$, to zmienna v będzie nieistotna dla struktury skupień, gdy $\phi_v = 1$, to zmienna ta będzie istotna. Idea znacznie upraszczająca postać (2.7) polega na tym, by zapisać (2.7) w postaci

$$f(x)=\sum_{k=1}^{K}\alpha_k\prod_{v=1}^{V}\left[\phi_v f_k\left(x_v|\theta_{kv}\right)+\left(1-\phi_v\right)q\left(x_v|\lambda_v\right)\right] \qquad (2.8)$$

jako, że ϕ_v jest zmienną binarną przyjmującą wartość 0 lub 1. Postać (2.8) pozwala na wyprowadzenie wersji wzorów z kroków E oraz M algorytmu EM. Metoda została zbadana na kilku zbiorach danych generowanych oraz na kilku empirycznych (trzy zbiory były z bazy UCI). Metoda spisała się dobrze na tych zbiorach.

W badaniu symulacyjnym przeprowadzonym przez Steinley'a i Brusco (2008), w oparciu o program komputerowy uzyskany od autorów, metoda wyrazistości cech wypadła dość dobrze uzyskując wskazania pamięci i precyzji zbliżone do najlepszych (91% i 81% odpowiednio). Nieco gorzej wypadła w świetle wskaźnika *ARI* (około 61%). Ogólnie, w grupie 8 metod zajęła 4-5 miejsce.

2.3. Metoda wyboru modelu

Metodę wyboru modelu zaproponowali Raftery i Dean (2006). Podobnie jak w przypadku metody wyrazistości cech zakładamy, że postać (2.2) jest modelem populacji przy czym ograniczamy się do skupień normalnych czyli $\theta_k = \left(\mu_k, \Sigma_k\right)$. Ponadto stosujemy rozkład spektralny macierzy kowariancji.

$$\Sigma_k = \lambda_k D_k A_k D_k \ , \qquad (2.9)$$

gdzie λ_k jest największą wartością własną macierzy kowariancji Σ_k; macierz D_k jest macierzą wektorów własnych macierzy kowariancji; A_k jest macierzą diagonalną zawierającą wartości własne na przekątnej. Te trzy czynniki mają za zadanie kontrolować objętość/liczebność skupienia, jego orientację i kształt. Przy pomocy takich trzech czynników można modelować zbiory, w których różne skupienia mogą mieć różne kształty (sferyczne, elipsoidalne, diagonalne) i liczebności. Następnie dzielimy wszystkie zmienne na trzy rozłączne podzbiory

V_1 - zbiór już wybranych zmiennych jako istotne;

V_2 - zbiór zmiennych kandydatek do włączenia do lub wykluczenia ze zbioru zmiennych istotnych;

V_3 - zbiór pozostałych zmiennych.

Decyzje o tym czy zbiór V_2 włączyć czy też nie do zbioru zmiennych istotnych podejmujemy w oparciu o porównanie dwóch modeli:

$$\begin{aligned} M^{(1)}: \quad & p(x|z) = p\left(V^{(3)}\middle|V^{(2)}, V^{(1)}\right) p\left(V^{(2)}\middle|V^{(1)}\right) p\left(V^{(1)}\middle|z\right) \\ M^{(2)}: \quad & p(x|z) = p\left(V^{(3)}\middle|V^{(2)}, V^{(1)}\right) p\left(V^{(2)}, V^{(1)}\middle|z\right) \end{aligned} \tag{2.10}$$

gdzie $p(\cdot|\cdot)$ jest symbolicznym zapisem wartości funkcji wiarygodności dla wcześniej oszacowanych parametrów. Model $M^{(1)}$ sugeruje, że zbiór zmiennych V_2 nie dostarcza dodatkowych informacji istotnych dla struktury skupień, zaś model $M^{(2)}$ sugeruje, że zbiór zmiennych V_2 dostarcza dodatkowe informacje w porównaniu ze zbiorem V_1.

W analizie skupień problem pogrupowania obiektów sprowadza się do wyboru odpowiedniego modelu reprezentującego inną kombinację liczby skupień. Ze zbioru wszystkich rozważanych modeli wybieramy ten, który ma najwyższą jakość dopasowania. Modele są wybierane przy pomocy kryterium *BIC* (*Bayesian information criterion*) postaci:

$$BIC = 2 \cdot \log\left(p\left(x|\hat{\theta}\right)\right) - l \cdot \log(n) \tag{2.11}$$

gdzie $\log\left(p\left(x|\hat{\theta}\right)\right)$ jest logarytmem funkcji wiarygodności dla oszacowanego wektora $\hat{\theta}$ parametrów modelu, zaś l jest liczbą parametrów tego modelu, którego wartość należy obliczyć dla wszystkich możliwych podzbiorów zbioru wszystkich zmiennych, wszystkich możliwych wartości K liczby skupień oraz 10 wyszczególnionych przez autorów tej metody postaci rozkładu spektralnego (2.9) macierzy kowariancji (por. tabela 15.1, *Gatnar i Walesiak,* 2009). Ponadto, należy ustalić algorytm kolejności wybierania zmiennych. Autorzy zaproponowali następujący algorytm:

Krok 1. Wybrać jako pierwszą zmienną tę, która ma najlepsze ze wszystkich zmiennych wskazania grupowalności.

Krok 2. Wybrać jako drugą zmienną tę, która ma wspólnie z pierwszą zmienną najlepsze wskazania dwuwymiarowej grupowalności dane wzorem (2.11).

Krok 3. Jako następną kandydatkę na zmienną istotną wziąć zmienną, która ma najlepsze wskazania wielowymiarowej grupowalności w połączeniu z

poprzednio wybranymi zmiennymi. Tę zmienną zaakceptować jeśli będą przeważały dowody o jej grupowalności.

Krok 4. Jako kandydatkę na zmienną nieistotną wziąć z bieżącego zbioru zmiennych wybranych tę, która ma najgorsze wskazania wielowymiarowej grupowalności w sensie porównania grupowalności wszystkich zmiennych dotychczas wybranych ze wszystkimi oprócz kandydatki. Tę zmienną zaakceptować (jako nieistotną) jeśli będą przeważały dowody o jej braku grupowalności.

Krok 5. Powtarzać kroki 3 i 4 do momentu aż nastąpią dwa kolejne odrzucenia.

Metoda została zbadana na kilku wygenerowanych zbiorach danych i na trzech zbiorach empirycznych (w tym dwa zbiory z bazy UCI). Na zbiorach tych metoda osiągnęła dobre wyniki.

W badaniu przeprowadzonym przez Steinley'a i Brusco (2008), w oparciu o program komputerowy autorstwa Martinez i Martinez (2001), omawiana metoda okazała się mało efektywna (pamięć - 33% , precyzja - 55% , *ARI* – 44%) pomimo założenia o znanej (i poprawnej) liczbie skupień i ujednoliceniu macierzy kowariancji dla wszystkich skupień. Wady przedstawionego podejścia całościowego tj. szacującego jednocześnie liczbę skupień, grupującego obserwacje i dokonującego selekcji zmiennych wymieniają też Fraiman i inni (2008).

2.4. Metoda separowalności rozrzutu

Metoda separowalności rozrzutu została opracowana przez Dy i Brodley'a (2004). Podobnie jak w poprzednich metodach autorzy modelują populację gęstością postaci (2.2) przy założeniu normalności skupień czyli $\theta_k = (\mu_k, \Sigma_k)$. Najpierw stosujemy algorytm EM do oszacowania parametrów modelu i do grupowania obserwacji. Następnie, stosujemy dwie miary rozproszenia

$$S_w = \sum_{k=1}^{K} \alpha_k \Sigma_k \quad \text{oraz} \quad S_b = \sum_{k=1}^{K} \alpha_k (\mu_k - M_0)(\mu_k - M_0)^T , \qquad (2.12)$$

gdzie $M_0 = \sum_{k=1}^{K} \alpha_k \mu_k$.

Miara S_w (macierz) mierzy rozrzut obserwacji względem środków ich skupień, zaś miara S_b (macierz) mierzy rozproszenie środków skupień μ_k względem średniej ogólnej czyli rozproszenie pomiędzy skupieniami. Miary te

są ważonymi wariantami miar (1.16) i (1.17) (por. § 1.2) w wersji dla populacji. Miara znaczenia dla struktury skupień dwóch porównywanych podzbiorów zmiennych *A* oraz *B* wykorzystuje wielkość

$$J(A, z_B) = \acute{s}lad\left(S_w^{-1} S_b\right). \tag{2.13}$$

którą należy rozumieć w tym sensie, że występującą po prawych stronach wzorów (2.14) stosujemy do wektora przynależności z_B (otrzymanego w oparciu o podzbiór *B* zmiennych) oraz tylko do zmiennych ze zbioru *A*. Ta miara jest obciążona w stosunku do wymiaru tzn. cechuje się wzrostem wartości wraz ze wzrostem liczby zmiennych w zbiorze *A* (nawet przy tym samym wektorze przynależności obserwacji). Miary znaczenia dla struktury skupień dwóch porównywanych podzbiorów zmiennych *A, B* dane są wzorami (2.14), które w umowny sposób normalizują miary (2.13) w stosunku do wektorów przynależności z_A oraz z_B:

$$\begin{aligned} ZnormMiara(A) &= J(A, z_A) J(B, z_A) \\ ZnormMiara(B) &= J(B, z_B) J(A, z_B) \end{aligned} \tag{2.14}$$

Poszukiwanie zbiorów zmiennych odbywa się przez kroczące poszukiwanie typu *forward search* przy czym zaczynać możemy od zbioru pustego. W każdym kroku dodajemy do zbioru zmiennych istotnych *A* zmienną *v*, jeżeli zwiększy ona wartość kryterium (2.14) tzn.

$$ZnormMiara(A \cup \{v\}) > ZnormMiara(A) \tag{2.15}$$

W badaniu przeprowadzonym przez Steinley'a i Brusco (2008) w oparciu o program komputerowy autorstwa Steinley'a i Brusco, metoda separowalności rozrzutu wypadła stosunkowo słabo (pamięć - 51% , precyzja - 61% , *ARI* – 39%) mając wskaźnik odzyskiwalności najsłabszy ze wszystkich ośmiu porównywanych metod.

2.5. Ocena podejść modelowych

Podsumowując metody modelowe wybierania zmiennych przedstawiające populację w postaci mieszaniny rozkładów normalnych należy stwierdzić, że zaletą tych metod jest ich ogólność polegająca na tym, że wraz z wnioskowaniem o liczbie zmiennych istotnych wnioskujemy o przypisaniu obiektów do skupień oraz o liczbie skupień. Metody te mają jednak istotne mankamenty, które nie pozwalają na pozytywną ocenę tej grupy metod. Mają one ograniczoną stosowalność w problemach praktycznych gdyż nadają się tylko do analizowania zbiorów, w których wszystkie zmienne opisujące obiekty są zmiennymi ciągłymi. Trzy metody z tej grupy przedstawione w tym rozdziale wypadły słabo w badaniu symulacyjnym (por. tabela 2.1), w którym zbiory

danych były z takich właśnie mieszanin generowane. Pozwala to wnioskować, że w przypadku empirycznych zbiorów danych, skupienia których na pewno nie zawsze są próbą z rozkładu normalnego (por. 5.1), efektywność tych metod będzie jeszcze niższa. Ponadto, gdy założenie o normalności skupień nie jest spełnione, to metody te tracą swoją najważniejszą zaletę – nie pozwalają na wyciąganie wniosków w formie wnioskowania statystycznego, są to tylko algorytmy postępowania. Inną, przyczyną słabej efektywności jest na pewno brak dobrego algorytmu szacowania parametrów modelu. Algorytm EM nie zawsze znajduje poprawne rozwiązanie globalne.

Tabela 2.1 Efektywność trzech metod modelowych dla zbiorów ze zmiennymi ciągłymi.

Metoda	Pamięć	Precyzja	*ARI*
Wyrazistości cech	0,911	0,813	0,607
Wyboru modelu	0,326	0,552	0,433
Separowalności rozrzutu	0,506	0,612	0,386

Źródło: Steinley i Brusco (2008).

W literaturze opinie o podejściach modelowych są raczej krytyczne (por. np. *Fraiman i inni,* 2008, *Steinley i Brusco,* 2008). Należy jednak zaznaczyć, że ocena efektywności wynikająca z tabeli 2.1 nie jest do końca obiektywna. Najlepsze z tych podejść – metoda wyrazistości cech, obok selekcji zmiennych szacuje również liczbę skupień na jakie należy pogrupować obiekty i dla takiej liczby była obliczana asymptotyczna odzyskiwalność *ARI*. Metody niemodelowe oceniane były przy zadaniu poprawnej liczby skupień, co było znaczącym ułatwieniem. Jeśli wziąć pod uwagę to zastrzeżenie, to metodę wyrazistości cech należałoby ocenić na równi z HINoV i jej dwiema modyfikacjami (por. roz. 3). Dwa pozostałe podejścia modelowe nie są godne uwagi – ich efektywność okazała się tak niska, że nawet jeśli wziąć pod uwagę wspomniane utrudnienia, trudno przypuszczać by okazały się użytecznymi w zastosowaniach do zbiorów empirycznych.

3. METODY HEURYSTYCZNE SELEKCJI ZMIENNYCH W ANALIZIE SKUPIEŃ

3.1. Uwagi wstępne

W odróżnieniu od podejścia modelowego przy konstruowaniu metod heurystycznych nie traktujemy zbioru danych jako próby wylosowanej z jakiejś populacji. Mamy do dyspozycji tylko macierz danych. Do chwili obecnej, jedną z najlepszych metod wybierania zmiennych w analizie skupień była metoda HINoV (*Carmone i inni,* 1999). Oceny innych metod które można znaleźć w literaturze stwierdzają, że metody wcześniejsze są słabsze lub dużo słabsze. W dalszym ciągu tego paragrafu przedstawione są krótkie charakterystyki metod opracowanych przed opublikowaniem HINoV.

Kruskal (1972*)* próbował znaleźć liniową transformację danych, która uwidoczniłaby strukturę skupień. Współczynniki takiej transformacji wyznaczane były przy pomocy indeksu skondensowania, który jest miarą spójności skupień. Zmienne bardziej istotne dla struktury skupień otrzymywały wyższe współczynniki. Nie udało mu się jednak znaleźć odpowiednio dobrego indeksu.

Jedną z pierwszych bardziej zaawansowanych metod wybierania zmiennych w analizie skupień był algorytm o nazwie SYNCLUS (*SYNthesised CLUStering* por. *DeSarbo i inni*, 1984). Algorytm ten miał na celu zarówno pogrupowanie obserwacji zmodyfikowaną metodą *k*-średnich jak i przypisanie wag zmiennym. Kryterium realizującym te zadania jest minimalizacja różnic pomiędzy ważonymi kwadratami odległości par obserwacji (przy zmiennych należących do a priori ustalonych grup zmiennych) oraz miary liczbowej oceniającej jakość grupowania. Wyrażenie, które należy zminimalizować to

$$Z = \sum_{i=1}^{I} \sum_{\substack{j,j'=1 \\ j<j'}}^{J} w_i^2 \left(\delta_{jj'} - d_{jj'}^{2(i)} \right)^2 \tag{3.1}$$

gdzie w_i^2 - waga (zakładamy, że znana a priori) przypisana *i*-tej grupie zmiennych ($\sum_i w_i^2 = 1$); $d_{jj'}^{2(i)}$ - kwadrat odległości pomiędzy obserwacjami $j,\ j'$ w *i*-tej grupie zmiennych dany wzorem

$$d_{jj'}^{2(i)} = \sum_{t_i=1}^{T_i} v_{it_i}^2 \left(x_{jt_i}^{(i)} - x_{j't_i}^{(i)} \right)^2 \tag{3.2}$$

zaś $\delta_{jj'} = \alpha\, a_{jj'}^{*} - \beta$; przy

$$a^*_{jj'} = \begin{cases} 1/n_k & \text{gdy obserwacje } j, j' \text{ należą do } k\text{ - tego skupienia} \\ 0 & \text{w przeciwnym przypadku} \end{cases} \tag{3.3}$$

a liczby α, β należy oszacować metodą najmniejszych kwadratów z J równań postaci $d^2_{jj'} = \alpha\, a^*_{jj'} - \beta$. We wzorze (3.2) $v^2_{it_i}$ oznacza wagę przypisaną zmiennej t_i w i-tej grupie zmiennych zaś $x^{(i)}_{jt_i}$ jest wartością zmiennej t_i dla j-ej obserwacji. Metoda ta została zbadana na kilku zbiorach danych empirycznych (*Green i inni,* 1990) i okazała się efektywna tylko na niektórych. Ponadto, stwierdzono, że metoda ma ograniczenia dla liczby zmiennych i liczby obserwacji. Niewątpliwie za ograniczenie należy też uznać konieczność podzielenia wszystkich zmiennych na grupy zmiennych logicznie połączonych, którym to grupom z góry nadajemy ustalone wagi.

Inną metodą wybierania zmiennych z tamtego okresu była metoda kroczącego wybierania zmiennych (*Fowlkes i inni*, 1988). Ideą tej metody jest zastosowanie hierarchicznego grupowania obserwacji a następnie dołączanie kroczące pojedynczych zmiennych zgodnie z kryterium wykorzystującym otrzymane grupowanie. Wskaźnikiem, który pokazuje, którą zmienną należałoby dołączyć do już wybranych jest wskaźnik oparty na wyborze zmiennej dającej jak największą wartość separowalności otrzymanego grupowania tzn.

$$S^*(k) = S(k) - E(S(k)) \tag{3.4}$$

gdzie

$$S(k) = \frac{1}{t}\sum_i \frac{e_i}{1+e_i} \tag{3.5}$$

gdzie e_i jest wartością własną macierzy $W^{-1}B$ (por. wzory 1.16-1.17), k jest liczbą skupień zaś $E(S(k))$ jest wartością oczekiwaną miary (3.5) , którą należy oszacować numerycznie przy założeniu, że nie ma wyraźnej struktury skupień, tj. na przykład, generując rozkłady wszystkich zmiennych potrzebnych do obliczenia $S(k)$ w postaci rozkładów normalnych, zaś $t = \min(V, k-1)$. Zauważmy, że istotną wadą takiej metody jest konieczność subiektywnej (na ogół wizualnej) interpretacji wartości kryterium (3.4). Autorzy zbadali tę metodę tylko na kilku przykładowych zbiorach danych (sztucznie wygenerowanych i empirycznych), jednak w obszerniejszym badaniu (*Donoghue,* 1995) okazało się, że metoda dała słabe wyniki w przypadku niektórych zbiorów. Trudno też nie zgodzić się z negatywną oceną tej metody (*Carmone i inni,* 1999) zarzucającą jej niemożność przeanalizowania wszystkich możliwych kombinacji

zmiennych dla dużych zbiorów danych, gdyż grupowanie hierarchiczne jest czasochłonne.

Inni badacze do identyfikacji zmiennych tworzących strukturę skupień próbowali wykorzystać statystyki oparte na skośności i kurtozie. Panuje ogólne przekonanie, że pojedyncze zmienne nie powinny mieć rozkładu normalnego, że powinien być on raczej skośny (por. *Morris i inni,* 1981, *Fletcher i Satz,* 1985). Bajgier i Aggarwal (1991) zbadali kilka wskaźników dla rozkładów jednowymiarowych i ustalili, że najlepsze były te oparte na ujemnej kurtozie rozkładu. Kierując się tymi doświadczeniami Donoghue (*Donoghue,* 1995) zaproponował wskaźnik postaci

$$m = g_2 - g_1^2 \tag{3.6}$$

gdzie g_1 jest współczynnikiem skośności

$$g_1 = \frac{M_3}{(M_2)^{3/2}} \tag{3.7}$$

zaś g_2 unormowaną kurtozą

$$g_2 = \frac{M_4}{(M_2)^2} - 3 \tag{3.8}$$

Autor proponuje wybieranie tylko tych zmiennych, dla których $m \le 0$ lub $m \le -1{,}2$. Drugim wskaźnikiem, który ten sam Autor proponuje jest wskaźnik wielomodalności

$$b = \frac{g_1^2 + 1}{g_2 + 3} \tag{3.9}$$

sugerując wybieranie zmiennych dla których $b > 0{,}555$ lub $b > 0{,}333$. Ponadto badano metodę opartą na samej tylko kurtozie tj. przy warununku $g_2 > -1{,}2$. W dość obszernym badaniu, w którym analizowano też metodę ważenia zmiennych dla grupowania przy pomocy drzew ultrametrycznych (por. *De Soete* 1986, 1988) oraz metodę kroczącego wybierania zmiennych (por. Fowlkes i inni, 1988), okazało się, że metody oparte na kurtozie spisały się słabo, metody oparte na współczynniku b oraz m wypadły słabo w przypadku występowania obserwacji nietypowych. Co więcej, w przypadku występowania obserwacji nietypowych słabo spisała się bardzo wymagająca pod względem obliczeniowym metoda kroczącego wybierania zmiennych.

De Soete (1986, 1988, por. *Makarenkov i Legendre,* 2001) zaproponował metodę optymalnego ważenia zmiennych, którą można wykorzystać dla różnych metod grupowania obserwacji. Ideą tej metody jest znalezienie wag $(w_1, w_2, \ldots, w_M)$, które będą minimalizowały funkcje straty postaci np.

$$L(w_1, w_2, \ldots, w_V) = \sum_{k=1}^{K} \frac{1}{n_k} \left(\sum_{i,j=1}^{n_k} d_{ij}^2 \right) \qquad (3.10)$$

dla grupowania metodą *k*-średnich, d_{ij}^2 oznacza kwadrat odległości ważonych wagami $w_1, w_2, \ldots, w_V$ obserwacji o numerach *i,j*, przy warunkach $w_i \geq 0$, $\sum_{i=1}^{V} w_i = 1$. Podobne funkcje straty są analizowane przez autorów (*Makarenkov i Legendre,* 2001) dla grupowania hierarchicznego przy pomocy drzew ultrametrycznych i addytywnych. Optymalne wagi znalezione przy pomocy numerycznych algorytmu optymalizacyjnego Polaka-Ribiere'go (por. np. *Press i inni,* 1986) można następnie użyć do obliczania odległości w procesie grupowania. Metoda ta została zbadana dość dokładnie (*Makarenkov i Legendre,* 2001) i okazała się nieodporna na obserwacje nietypowe. Autorzy doszli nawet do wniosku, że w przypadku występowania obserwacji nietypowych można użyć dwojakiego rodzaju ważenia, a mianowicie, najpierw ważyć obserwacje a potem zmienne. Takie propozycje należy uznać za mało praktyczne, gdyż, tak naprawdę, nie ma wiarygodnego sposobu rozstrzygania kwestii czy obserwacje nietypowe istnieją czy nie i nie wiadomo nic na temat ewentualnej efektywności takiej metody. Należy zatem zgodzić się z negatywną oceną tej metody zamieszczoną w pracy *Carmone i inni* (1999).

Metodą trochę podobną do metody DeSarbo była metoda zaproponowana przez Van Buurena i Heisera (1989). Metoda ta jest ogólniejsza w tym sensie, że nie wymaga podzielenia zmiennych na znane a priori, logicznie połączone grupy zmiennych. Natomiast jest znacznie mniej ogólna w tym sensie, że nadaje się tylko do zmiennych mierzonych na słabych skalach (nominalnych i porządkowych). Polega ona na tym by jednocześnie pogrupować obserwacje zgodnie z kryterium podobnym do kryterium z metody *k*-średnich, a następnie dokonać skalowania zmiennych optymalnego (por. analiza korespondencji np. *Gatnar, Walesiak,* 2004) dla otrzymanego grupowania. Dokładniej, można te metodę krótko scharakteryzować w następujący sposób. Niech $\boldsymbol{k} = (k_1, \ldots, k_V)$ będzie wektorem opisującym liczbę kategorii każdej ze zmiennych opisujących obserwacje zbioru danych; G_j niech będzie macierzą wskaźnikową *j*-ej zmiennej tj. macierzą, która w *i*-tym wierszu odpowiadającym *i*-tej obserwacji ma jeden niezerowy element wskazujący na to, który wariant na *j*-ej zmiennej przyjęła *i*-ta obserwacja; $X_{n \times p}$ niech będzie nieznaną macierzą rzutowań obserwacji *V*-wymiarowych na przestrzeń *p*-wymiarową; Y_j niech będzie nieznaną macierzą o wymiarach $k_j \times p$ rzutowań wszystkich kategorii *j*-ej

zmiennej na tę samą przestrzeń p-wymiarową. Zadanie wybierania zmiennych realizujemy w tej metodzie przez znalezienie jak najlepiej dobranych p oraz macierzy $X_{n\times p}$ i Y_j co osiągamy minimalizując wartość wyrażenia

$$\frac{1}{V}\sum_{j=1}^{V} tr\left(X - G_j Y_j\right)^T \left(X - G_j Y_j\right) \qquad (3.11)$$

Minimalizując wartość wyrażenia (3.11) minimalizujemy sumę kwadratów odległości, która to metoda, jak sami autorzy przyznają, obarczona jest znaną wadą metody k-średnich polegającą na słabej efektywności przy różnych liczebnościach skupień i przy ich nieregularnym (np. łańcuchowym) kształcie. Ponadto, w przykładach zastosowań prezentowanych przez nich widać duży subiektywizm w możliwości wyboru p oraz liczby skupień k – sugerują nawet posłużenie się zjawiskiem łokcia przy wyborze k. Na tym (czyli początkowym) etapie posługiwanie się tak mało precyzyjnym kryterium może budzić wątpliwości co do efektów dokonanych wyborów. Wymienione wady upoważniają do tego by można było zgodzić się z negatywną oceną możliwości szerszych zastosowań tej metody zawartą w pracy Carmone (*Carmone i inni,* 1999) gdzie oceniono ją jako nieefektywną z powodu często zdarzającej się niemożności uzyskania optymalnego rozwiązania globalnego oraz z powodu komplikacji natury obliczeniowej.

Przy wybieraniu zmiennych w analizie skupień naturalnym podejściem jest próba scharakteryzowania pojedynczych zmiennych przy pomocy określonych wskaźników, które pomogłyby w rozstrzygnięciu kwestii czy dana zmienna jest istotna dla struktury skupień. Jak zaznaczono w § 1.3 badanie charakterystyk pojedynczych zmiennych, na ogół, jest mało użyteczne. Ponadto, intuicyjnie rzecz biorąc jakość takich wskaźników będzie się pogarszała wraz ze wzrostem liczby zmiennych i liczby skupień, tym nie mniej, próby konstruowania takich wskaźników są podejmowane przez wielu badaczy. Można wskazać dwie sytuacje, w których wskaźniki zdolności dyskryminacyjnej pojedynczych zmiennych są użyteczne. Pierwsza to ta, w której struktura skupień jest tak silna, że do zaobserwowania nawet na rozkładach brzegowych. Druga sytuacja może zaistnieć, na przykład wtedy, gdy badamy zbiór zmiennych bardzo silnie skorelowanych. Wówczas, paradoksalnie, analizowanie zdolności dyskryminacyjnej podzbiorów tego zbioru wnosi niewiele nowego do tego co uzyskamy z analizy pojedynczych zmiennych i wskaźniki je charakteryzujące mogą znaleźć zastosowanie.

Propozycją mierzenia skupialności czy też zdolności dyskryminacyjnej (zdolności grupowania) jednej zmiennej może być wskaźnik postaci (*Sokołowski,* 1992)

$$G_v = 1 - \frac{1}{r(v)} \sum_{i=1}^{n-1} \min \left\{ x_{v(i+1)} - x_{v(i)} \; ; \; \frac{r(v)}{n-1} \right\} \qquad (3.12)$$

pod warunkiem, że $r(v) > 0$ ($x_{v(i)}$ - uporządkowane niemalejąco wartości v-ej zmiennej). Z postaci wzoru (3.12) wynika, że jest on próbą ocenienia czy pojedyncza zmienna ma rozkład obserwacji charakteryzujący się występowaniem obszarów o większym zagęszczeniu obserwacji. Przykład zamieszczony w pracy Walesiaka (2005) pokazuje, że można tak dobrać wartości dla dwóch zmiennych o rozkładach równomiernych (czyli o wartościach współczynnika (3.12) równych 0), że mimo to będą tworzyły razem tj. w dwóch wymiarach wyraźną strukturę skupień. Sokołowski uogólnił wskaźnik (3.12) na przypadek wielowymiarowy, ale uogólnienie to jest uzależnione od metody najbliższego. Uogólnienie ma postać :

$$S = 1 - \frac{1}{\sum_{i=1}^{m-1} a_i} \sum_{i=1}^{m-1} \min \left\{ a_i \; ; \; \frac{\sum_{i=1}^{m-1} a_i}{m-1} \right\} \qquad (3.13)$$

gdzie a_i - i-te wiązadło w dendrycie; i=1,...,m-1 numer wiązadła; m – liczba wierzchołków w dendrycie. Statystyka (3.13) przyjmuje wartości z przedziału $[0;1-1/(n-1)]$ i mierzy „odległość" pomiędzy strukturą dendrytu empirycznego i teoretycznego. Im wartości bliższe 1 tym większe możliwości wykrycia struktury skupień przy pomocy metody najbliższego sąsiada. Wskaźnik (3.12) dość dobrze odróżnia rozkłady istotne dla struktury skupień od rozkładów równomiernych ale przyjmuje bardzo podobne wartości dla rozkładów tworzących strukturę skupień i dla niektórych rozkładów jednomodalnych. Fakt ten jest spowodowany tym, że wskaźnik nie wychwytuje lokalnych różnic w gęstości rozkładu zmiennej a jedynie sprawdza czy występują obszary o większej gęstości.

Dalszy ciąg rozdziału zawiera scharakteryzowanych dokładniej kilka najnowszych metod wybierania zmiennych istotnych. Najlepsze spośród nich są oceniane w eksperymencie symulacyjnym.

3.2. Metoda Brusco dla zmiennych binarnych

Wybieranie zmiennych istotnych dla struktury skupień wtedy gdy zbiór danych opisany jest tylko przez zmienne mierzone na skali nominalnej (w szczególności binarnej) jest zadaniem specyficznym, istotnie różniącym się od takiego samego zadania dla zbioru ze zmiennymi mierzonymi na silnych skalach. Specyfika tego zadania polega na tym, że, formalnie rzecz biorąc, wszystkie zmienne nominalne można uznać za istotne ustalając tak dużą liczbę

skupień jaka wynikałaby z liczby maksymalnych liczebnościowo zbiorów obiektów opisanych przez kombinacje wartości wszystkich zmiennych opisujących obiekty. W przypadku silnych skal pomiarowych taka liczba skupień byłaby na ogół równa liczebności zbioru danych, co jest jawną sprzecznością z ideą analizy skupień. Ale w przypadku zmiennych binarnych, liczba skupień wynikająca ze wspomnianego warunku byłaby godna rozważenia, zwłaszcza przy małej liczbie zmiennych. Taka liczba skupień, w przypadku wielokrotnego powtarzania się układów takich samych wartości wszystkich zmiennych byłaby znacznie mniejsza od liczby obiektów w zbiorze danych, co pozostawałoby w zgodności z celem analizy skupień. Konstruując metody wybierania zmiennych dla zbiorów danych ze zmiennymi wyłącznie nominalnymi należy jednak ustosunkować się do zagadnienia liczby skupień (por. też 4.4.2). Brusco (2004) rekomenduje najpierw użycie jakiejś dobrej metody szacowania liczby skupień w zbiorze danych binarnych (np. *Ratkowsky i Lance,* 1978), a następnie, znając liczbę skupień, optymalizację doboru zmiennych przy wykorzystaniu metody grupowania *k*-średnich. Funkcją dyskryminacyjną decydującą o wyniku optymalizacji jest funkcja

$$Z = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{(i<j)\in C_k} d_{ij} \qquad (3.14)$$

którą minimalizujemy po wszystkich podziałach $\pi = \{C_1, C_2, \ldots, C_k\}$ zbioru wszystkich rozważanych obiektów (na ogół próby z całego zbioru danych). Miarą odległości jest metryka

$$d_{ij} = \left|\left(S_i \cap \overline{S}_j\right) \cup \left(S_j \cap \overline{S}_i\right)\right| \qquad (3.15)$$

gdzie S_i oznacza zbiór zmiennych mających wartość równą 1 dla *i*-tego obiektu zaś $\overline{S}_i$ dopełnienie tego zbioru do wszystkich zmiennych czyli zbiór zmiennych mających wartość równą 0 dla *i*-tego obiektu. Wzór ten jest tożsamy z formułą Sokala-Michenera (por. tabela 1.1). Kolejne kroki algorytmu wyglądają następująco:

Krok 1. Rozważamy podzbiór *L* (o liczebności *M*) zbioru wszystkich *n* obiektów wylosowany ze zbioru danych. Kładziemy $Z_{\min} = \infty$ oraz $Z_{B2} = \infty$.

Krok 2. Dla każdego z $\binom{V}{P}$ podzbioru *P*-elementowego zbioru wszystkich zmiennych utwórzmy macierz *A* opisującą wylosowane obiekty. Oznaczmy zbiór zmiennych reprezentujących kolumny macierzy *A* symbolem Q'.

Krok 3. Wykonujemy 500 grupowań metodą *k*-średnich obiektów zbioru L i przez $Z_{\min}$ oznaczamy minimalną wartość wyrażenia (3.14).

Krok 4. Próbujemy dołączać każdą pojedynczą zmienną z pozostałych zmiennych $p \in VV \setminus Q$ przy pomocy 5000 powtórzeń grupowania metodą *k*-średnich elementów zbioru L dla zmiennych $Q \cup \{p\}$. Oznaczmy przez $Z_{\min}$ minimalną wartość wyrażenia (3.14), zaś zmienną odpowiadającą temu wyrażeniu przez p'.

Krok 5. Jeśli $Z_{\min} < Z_{B2}$, to kładziemy $Z_{B2} = Z_{\min}$ i akceptujemy zmienną *p*.

Krok 6. Jeśli $Z_{B2} > Z_{\min} + \delta \frac{M}{4}$, to kończymy proces selekcji. W przeciwnym razie kładziemy $Q = Q \cup \{p'\}$ i idziemy do kroku 4.

Efektywność metody została zbadana dla $\delta = 0{,}5$. Metoda spisała się bardzo dobrze w eksperymencie symulacyjnym, w którym parametrami były:

- liczba obiektów (2000, 3000, 4000);
- liczba skupień (4, 6, 8);
- liczba zmiennych istotnych (4, 6, 8);
- liczebności skupień (od równych do umiarkowanie zróżnicowanych);
- poziom zanieczyszczeń (błędów losowych) (0, 2%, 4%);
- liczba zmiennych maskujących (0, 4, 8).

Należy zauważyć, że wadą takiego eksperymentu jest zakładanie znanej poprawnej liczby skupień. Stwierdzenie autora, że istnieją doskonałe metody szacowania liczby skupień w zbiorze danych nie ma w tym przypadku żadnego sensu. Przypadek ten nie różni się bowiem niczym od przypadku ogólnego tzn. dowolnych skal pomiarowych, w którym, przed wybraniem zmiennych szacowanie liczby skupień raczej nie ma żadnego sensu. Dość duże błędy przy szacowaniu liczby skupień zdarzają się nawet przy bezbłędnie określonym zbiorze zmiennych tworzących strukturę skupień. Ponadto sposób definiowania zmiennych nieistotnych jest w przypadku zmiennych nominalnych kwestią dość umowną, mocno uzależnioną od liczby obiektów zbioru. Gdy w grę wchodzą zmienne wielostanowe, to kwestia ta jest dość trudna do rozstrzygnięcia, gdyż nie wiadomo jakie są granice tolerancji dla stopnia losowości przy ustalaniu wartości zmiennych nieistotnych dla poszczególnych obiektów. Kwestia doboru wartości zmiennych nieistotnych jest odpowiednikiem doboru typu rozkładu nieistotnego dla struktury skupień w przypadku zmiennych ciągłych. Bardzo umowny charakter ma też dobór stopnia poziomu zanieczyszczeń, który jest odpowiednikiem stopnia zachodzenia skupień na siebie w przypadku zmiennych

ciągłych. Poziom ten został obniżony (do 3%) w drugim eksperymencie, którym badano efektywność metody w przypadku istnienia wielokrotnych struktur skupień i, co za tym idzie, nieco większej liczby zmiennych. W tym eksperymencie metoda zdołała zawsze wybrać jeden ze zbiorów zmiennych opisujących jedną ze struktur. Jednak powodem obniżenia stopnia rozmycia skupień był zapewne fakt pojawienia się większej (18) liczby zmiennych.

3.3. Metoda Fishera-Talavery dla zmiennych nominalnych

Metodę wybierania zmiennych przeznaczoną dla zmiennych mierzonych na skali nominalnej zaproponowali Talavera i Fisher (*Talavera,* 2000, *Fisher i Talavera*, 1987). Ideą tej metody jest eliminowanie zmiennych wykazujących niską korelację z pozostałymi zmiennymi. Korelację pomiędzy zmiennymi mierzymy w następujący sposób. Jeżeli zbiór danych został podzielony na skupienia jak najmniej zróżnicowane wewnętrznie i jak najbardziej zróżnicowane między sobą, to zarówno odsetek obserwacji z *k*-tego skupienia, dla których zmienna o numerze *v* przyjmie wartość a_{vj} (czyli swój *j*-y wariant), który jest prawdopodobieństwem $P(C_k | x_v = a_{vj})$, jak i odsetek wartości zmiennej o numerze v , które są równe a_{vj} dla obserwacji z *k*-tego skupienia, który zapiszemy w postaci $P(x_v = a_{vj} | C_k)$ powinny mieć wysokie wartości. Wobec tego jakość podziału zbioru możemy mierzyć wartością wyrażenia

$$\sum_{k=1}^{K} \sum_{v} \sum_{j} P(x_v = a_{vj}) P(C_k | x_v = a_{vj}) P(x_v = a_{vj} | C_k) \qquad (3.16)$$

We wzorze tym $P(x_v = a_{vj})$ oznacza odsetek obserwacji a_{vj} wśród wszystkich wartości zmiennej o numerze *v* i odgrywa rolę wagi przypisanej iloczynowi prawdopodobieństw decydujących o jakości podziału. Korzystając z twierdzenia Bayesa możemy zapisać

$$P(x_v = a_{vj}) P(C_k | x_v = a_{vj}) = P(C_k) P(x_v = a_{vj} | C_k) \qquad (3.17)$$

i podstawiając do (3.16) otrzymamy, że jakość podziału możemy mierzyć za pomocą

$$\sum_{k=1}^{K} P(C_k) \sum_{v} \sum_{j} P^2 (x_v = a_{vj} | C_k) \qquad (3.18)$$

Podwójną sumę wewnętrzną możemy traktować jako wartość średnią liczby prawidłowo odgadniętych wartości wszystkich cech dla dowolnego obiektu z klasy C_k . Taka interpretacja jest dozwolona jeśli założymy, że wartość

dowolnej zmiennej odgadujemy z prawdopodobieństwem $P\left(x_v = a_{vj} \middle| C_k\right)$ i, że wartość ta jest przyjmowana przez zmienną z takim samym prawdopodobieństwem. Przy takiej interpretacji wartość oczekiwana liczby prawidłowo odgadniętych wartości zmiennych bez odwoływania się do podziału zbioru na skupienia jest równa $\sum_v \sum_j P^2\left(x_v = a_{vj}\right)$. Odejmując tę sumę od sumy wewnętrznej ze wzoru (3.18) otrzymamy przyrost wartości oczekiwanej liczby prawidłowo odgadniętych wartości zmiennych jaki daje nam znajomość podziału zbioru obiektów na skupienia. Wzór (3.18) możemy łatwo zaadoptować do skonstruowania formuły, która określi zależność ustalonej cechy v_M od wartości przyjmowanych przez pozostałe cechy, podstawiając $P\left(x_v = a_{vj}\right)$ zamiast $P\left(C_k\right)$ i zamieniając sumowanie po wszystkich skupieniach *k* na sumowanie po wszystkich zmiennych *v* i ich wariantach *j*. Wykorzystując to odejmowanie prawdopodobieństw i uśredniając wynik otrzymamy wyrażenie

$$Kor\left(v_M\right) = \frac{\sum_v \sum_j P\left(x_v = a_{vj}\right) \sum_{j_M} \left(P^2\left(x_{v_M} = a_{vj_M} \middle| x_v = a_{vj}\right) - P^2\left(x_{v_M} = a_{vj_M}\right)\right)}{\left|\left\{v \middle| v \neq v_M\right\}\right|} \tag{3.19}$$

gdzie wielkość z mianownika to liczba cech różnych od cechy v_M .

Wyrażenie (3.19) można wykorzystać do uporządkowania cech od cechy najsilniej skorelowanej z pozostałymi cechami do cechy skorelowanej najsłabiej. W przypadku zbioru zmiennych ograniczonego tylko do tych mierzonych na skali nominalnej takie uporządkowanie można, przy założeniu poczynionym przez autorów, utożsamiać z hierarchią istotności cech dla struktury skupień. Chcąc wykorzystać ustalony porządek cech do wyboru cech istotnych możemy w dalszej kolejności zastosować określone kryterium podziału na dwa zbiory zależne od ustalonego algorytmu grupowania obserwacji (będzie to metoda typu *wrapper*) lub niezależne od żadnego algorytmu grupowania (będzie to metoda typu *filter*).

Autorzy zastosowali następujący sposób ocenienia efektywności metody. Użyto algorytmu COBWEB (*Fisher,* 1987) opartego na drzewie hierarchicznym do grupowania obiektów zbiorów danych. Rozważyli kilka zbiorów danych, dwa sztuczne i sześć z repozytorium UCI Uniwersytetu Kalifornijskiego. Każdy zbiór podzielony został losowo na dwa równe podzbiory: treningowy i testujący. Następnie do zbioru treningowego zastosowano algorytm COBWEB i wyniki

grupowania użyte zostały do grupowania obserwacji ze zbioru testowego. Wyniki grupowania oceniano przy pomocy testu opartego na sprawdzaniu krzyżowym (*Dietterich,* 1998). Dla wybranych *k* zmiennych najistotniejszych w hipotezie stwierdzającej, że użycie wszystkich zmiennych daje mniejszy odsetek błędnych klasyfikacji od wybranych *k* zmiennych, można, jak dowodzi autor tego artykułu, pokładać bardzo małe zaufanie. Miernikiem liczbowym dla wybierania *k* jest najmniejszy odsetek błędów klasyfikacji (w porównaniu zbioru testowego z uczącym). Jednak na niektórych spośród sześciu prezentowanych wykresów jest wiele kandydatów do tej roli (odsetki błędów są bardzo zbliżone do siebie).

Zauważmy, że dzięki założeniu sformułowanym w punkcie wyjścia (o utożsamieniu zmiennych istotnych ze zmiennymi najsilniej skorelowanymi ze wszystkimi pozostałymi zmiennymi) omijamy problem odniesienia się do liczby skupień. To jest niewątpliwie zaletą metody, choć do końca tego problemu nie da się takim sposobem ominąć (por. 4.4.2). Wprowadzone przez autorów założenie ma dobre konsekwencje gdy liczba zmiennych istotnych wśród wszystkich kandydatek jest znacząca tj. na przykład, większa od połowy. Gdy zmiennych tworzących strukturę skupień jest niewiele w stosunku do liczby wszystkich zmiennych, to intuicyjnie rzecz biorąc, założenie o konieczności skorelowania zmiennych istotnych ze wszystkimi zmiennymi traci sens wraz ze spadkiem udziału liczby zmiennych istotnych we wszystkich zmiennych. Można wyobrazić sobie sytuację, w której kilka zmiennych nominalnych tworzy wyraźną strukturę skupień, ale oprócz tych zmiennych istnieje druga grupa zmiennych o rozkładzie równomiernym swoich stanów. Wraz ze wzrostem skorelowania w drugiej grupie zmiennych zaczynają one tworzyć drugą strukturę skupień i o rezultacie wybierania zmiennych zaczynają decydować siły związków pomiędzy zmiennymi i wzajemne liczebności obu grup. Wobec tego, metoda oparta zatem na wybieraniu zmiennych najsilniej skorelowanych ze wszystkimi zmiennymi nie nadaje się do przypadku występowania wielokrotnych struktur skupień.

3.4. Metoda COSA

Metoda COSA (*Clustering Objects on Substes of Attributes*) (por. *Friedman i Meulman,* 2004) polega na znalezieniu dla każdej zmiennej wagi uzależnionej od przekształconej wykładniczo odległości pomiędzy obserwacjami oraz od zastosowanego algorytmu grupowania obserwacji. Oryginalna cecha tej metody polega na tym, że dla różnych skupień dopuszczamy możliwość tworzenia struktury skupień przez różne podzbiory zmiennych. Wobec tego, w najogólniejszym przypadku zakładamy, że zbiorowi wszystkich zmiennych

odpowiada macierz W, która jest macierzą wag w_{vk} zmiennych w poszczególnych skupieniach czyli w_{vk} to waga *v*-ej zmiennej w *k*-ym skupieniu; W_k (kolumna macierzy *W*) to wektor wag zmiennych w *k*-ym skupieniu. Takie podejście jest obce innym metodom, wobec tego, w celu porównania tej metody z innymi metodami w eksperymencie symulacyjnym należy założyć jednolitość zbioru zmiennych dla wszystkich skupień. W badaniu przeprowadzonym przez Steinley'a i Brusco (2008) autorzy zaadoptowali tę metodę do takiego właśnie przypadku. Ponadto, ten uproszczony problem selekcji zmiennych wymaga tego by, wszystkie skupienia były opisane przez te same zmienne wobec czego macierz W można przekształcić do wektora o V współrzędnych, czyli do wektora składającego się z wag zmiennych. Kolejne kroki zastosowanego przez nich algorytmu metody są następujące.

Krok 1. Połóż $W_{V\times K} = \left\{\frac{1}{V}\right\}$; $\eta = \lambda$.

Krok 2. Początek pętli :

$$D_{ij}^{(\eta)}[W] = -\eta \log \sum_{v=1}^{V} w_v e^{-d_{ijv}/\eta}, \qquad (3.20)$$

gdzie $d_{ijv} = \left(x_{iv} - x_{jv}\right)^2 / s_v^2$, oraz $s_v = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_{iv} - x_{jv}\right)^2$ przy warunkach $w_v \geq 0$ oraz $\sum_{v=1}^{V} w_v = 1$.

Krok 3. Zastosuj ustalony algorytm grupowania obserwacji używając odległości pomiędzy obserwacjami takiej jak w kroku drugim.

Krok 4. Oblicz wagę *v*-ej zmiennej w *k*-tym skupieniu

$$w_{vk} = \frac{\exp\left(-s_{vk}/\lambda\right)}{\sum_{v'=1}^{V}\exp\left(-s_{v'k}/\lambda\right)} \qquad (3.21)$$

gdzie $s_{vk} = \frac{1}{n_k^2}\sum_{i\in C_k}\sum_{j\in C_k} d_{ijv}$, zaś wagi ostateczne *v*-ej zmiennej za pomocą wzoru

$$w_v = \sum_{k=1}^{K}\frac{n_k w_{vk}}{n} \qquad (3.22)$$

Krok 5. Zakończ pętlę gdy macierz W ustabilizuje swoje wartości. W przeciwnym przypadku połóżmy $\eta = \eta + \delta\lambda$ i wróćmy do początku pętli.

W wspomnianym badaniu (*Steinley i Brusco,* 2008), w kroku 3 algorytmu jako metodę grupowania zastosowano procedurę aglomeracyjną średniego połączenia. Natomiast parametry λ oraz δ zostały ustalone jako $\lambda = 0{,}2$ i $\delta = 0{,}1$. Metoda COSA okazała się średnio efektywna uzyskując wyniki : pamięć - 66% , precyzja - 74% , *ARI* – 72%.

Należy zaznaczyć, że cytowana ocena metody w żadnym razie nie może być traktowana jako ostateczna. Oceniona została bowiem tylko podstawowa, najwęższa wersja metody. Oryginalna pełna wersja skonstruowana została pod kątem rozwiązywania problemu selekcji zmiennych daleko ogólniejszego od zawężonej formy problemu rozpatrywanego przez Steinley'a i Brusco (2008). Wagi (3.21) zmiennej v w poszczególnych skupieniach są bowiem wagami minimalizującymi wartość przyjętego kryterium

$$Q(c,W) = \sum_{k=1}^{K} \frac{W_k}{n_k^2} \sum_{c(i)=l} \sum_{c(j)=l} D_{ij}^{(\lambda)}(W_k), \qquad (3.23)$$

w którym, nie stosując ujednolicenia wag zmiennych we wszystkich skupieniach, można, nadając im wagi uzależnione od skupienia, rozpatrywać wielokrotne struktury skupień dopuszczając ewentualność powtarzania się niektórych zmiennych w różnych strukturach. Parametr λ (*tuning parameter*) w równaniu (3.24) (i co za tym idzie również (3.23)) kontroluje liczebność zmiennych w skupieniach tzn. im większa wartość λ, tym więcej zmiennych może mieć wysoką wagę w danym skupieniu. Kryterium (3.23) jest modyfikacją kryterium wyjściowego (3.24) nie zawierającą parametru λ. Kontrolowanie wpływu λ na odległości pomiędzy obiektami zbioru dane jest formułą:

$$D_{ij}^{(\lambda)}(W_k) = \sum_{v=1}^{V} \left(w_{vk} d_{ijv} + \lambda w_{vk} \log(w_{vk}) \right) + \lambda \log(V), \qquad (3.24)$$

której idea polega na tym by dołączyć do odległości wyrażenie $\lambda w_{vk} \log(w_{vk})$, które ma być karą za wykorzystywanie do opisania struktury skupień większej liczby zmiennych. Przy ustalonej funkcji c przypisującej obiekty do skupień optymalne wagi tj. wagi minimalizujące kryterium (3.23) dane są równaniem (3.21). W celu usprawnienia algorytmu poszukiwania optymalnego rozwiązania posłużono się miarą odległości (występujących we wzorze (3.20)), która jest średnią ważonych odwrotności wykładniczych odległości pierwotnych. Ze wszystkich koncepcji selekcji zmiennych w analizie skupień metoda COSA jest najogólniejszą. Przeciętna efektywność jaką metoda ta uzyskała w najbardziej typowej, ograniczonej formie problemu selekcji zmiennych jest zapewne ceną jaką metoda płaci za swą ogólność.

3.5. Metoda kolejnych rzutowań

Metoda kolejnych rzutowań (*projection pursuit*) zbioru danych ma różne zastosowania w statystyce. Do wybierania zmiennych w analizie skupień zastosowały to podejście Montanari i Lizzani (2001). Metoda ta polega na tym, że do macierzy danych $\boldsymbol{X}$ stosujemy rzutowania jednowymiarowe i w wyniku rzutowań otrzymujemy n-wymiarowy wektor

$$y(a) = aX^T \qquad (3.25)$$

gdzie V-wymiarowy wektor poziomy $\boldsymbol{a}$ reprezentuje „kierunek" rzutowania. Współrzędne wektora $\boldsymbol{a}$ posłużą do konstrukcji wag jakie będą przypisane zmiennym, wobec czego nakładamy warunek by wektor ten m iał długość równą 1. Ponadto wektor $\boldsymbol{a}$ powinien być tak wyznaczony, by wektor $\boldsymbol{y}(\boldsymbol{a})$ miał postać jak najbardziej oddaloną od wektora o współrzędnych normalnych. Oddalenie będziemy mierzyć przy pomocy indeksu rzutowania Indeksem rzutowania może być dowolna funkcja, która osiąga minimum dla rozkładu normalnego np. ujemna entropia Shannona, informacja Fishera, statystyka Kołmogorowa-Smirnowa, statystyka χ^2. Autorki zdecydowały się wybrać statystykę χ^2 z powodu jej atrakcyjności obliczeniowej. A zatem, wektor $\boldsymbol{a}$ powinien być tak wyznaczony by zmaksymalizować wartość $\chi^2(\boldsymbol{y}(\boldsymbol{a}))$ przy warunku $\boldsymbol{a}\boldsymbol{a}^{\boldsymbol{T}} = 1$. Do rozwiązania tego problemu optymalizacyjnego stosuje się procedurę symulacyjnego wypalania zmiennych (*simulated annealing*) (por. np. *Goffe i inni*, 1994). Wagę która będzie przypisana i-ej zmiennej, które będą decydować o wyborze zmiennych, oblicza się za pomocą współczynnika istotności (*importance coefficient*)

$$CI_i = \frac{|a_i| s_i}{\sqrt{\sum_{j=1}^{V} (a_j s_j)^2}} \quad . \qquad (3.26)$$

Następnie ustalamy próg dla wartości współczynnika istotności poniżej którego zmienne mające odpowiadający im współczynnik uznamy za nieistotne. Progiem tym jest 10-ty percentyl równomiernego rozkładu brzegowego współczynnika istotności. Po wyrzuceniu zmiennych o współczynnikach poniżej tego progu powtarzamy procedurę symulacyjnego wypalania zmiennych na zmniejszonym zbiorze zmiennych, do momentu aż żadna zmienna nie zostanie wyrzucona.

W badaniu symulacyjnym przeprowadzonym przez Steinley'a i Brusco (2008) metoda kolejnych rzutowań uplasowała się na czwartym miejscu (pamięć - 73% , precyzja - 71% , *ARI* – 70%).

3.6. Metoda HINoV

Metoda HINoV została zaproponowana w roku 1999 (*Carmone i inni*). Była to dość przełomowa propozycja dla zagadnienia wybierania zmiennych w analizie skupień, gdyż później nikt już raczej nie wracał do metod wcześniejszych, które we wspomnianej publikacji zostały skrytykowane. Metoda HINoV w wersji dla grupowania metodą *k*-średnich składa się z następujących kroków.

Dla każdej zmiennej *v* przeprowadzamy grupowanie metodą *k*-średnich. Następnie, obliczamy skorygowany indeks Randa dla każdej pary zmiennych, który zgodnie ze wzorem (1.12) przyjmie postać:

$$RI(u,v)=\frac{\binom{n}{2}(t_1+t_2)-[(t_1+t_3)(t_1+t_4)+(t_3+t_2)(t_4+t_2)]}{\binom{n}{2}^2-[(t_1+t_3)(t_1+t_4)+(t_3+t_2)(t_4+t_2)]} \tag{3.27}$$

gdzie

t_1 - liczba par obserwacji należących do tego samego skupienia dla obu zmiennych;

t_2 - liczba par obserwacji należących do innych skupień dla obu zmiennych;

t_3 - liczba par obserwacji należących do tego samego skupienia dla zmiennej *u* i do różnych skupień dla zmiennej *v;* t_4 - liczba symetryczna do t_3.

Wartość indeksu $RI(u,v)$ interpretuje się jako miarę podobieństwa dwóch podziałów tego samego zbioru obserwacji, przy czym każdy podział dokonany jest w oparciu o jedną ze zmiennych *u, v*. Następnie, dla każdej zmiennej obliczamy sumę

$$TOPRI(u)=\sum_{u\neq v} RI(u,v), \tag{3.28}$$

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 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 $TOPRI(u)$, który można wizualnie zbadać na wykresie osypiska (na

osi rzędnych wskaźnik $TOPRI(u)$, na osi odciętych u, por. rys. 1.3). Zmienne z małymi wartościami $TOPRI(u)$ to zmienne nieistotne. Ta cecha konieczności subiektywnej, wizualnej interpretacji zastosowanego wskaźnika jest zasadniczą wadą tej metody.

W badaniu symulacyjnym (*Steinley i Brusco,* 2008) autorzy nie mogli wizualnie oceniać każdego wykresu osypiska wobec czego posłużyli się kryterium największego skoku wskaźnika $TOPRI(u)$. Kryterium polega na tym, ze po uporządkowaniu wartości tego wskaźnika malejąco obliczamy ilorazy

$$RR(s) = \frac{TOPRI(k(s)) - TOPRI(k(s+1))}{TOPRI(k(s-1)) - TOPRI(k(s))} \tag{3.29}$$

i wybieramy s początkowych zmiennych do największej wartości $RR(s)$. Stosowanie takiego kryterium (największego skoku) nie jest pozbawione wad. Z jednej strony jest ułatwieniem zadania wybierania zmiennych, gdyż z góry zakładamy, że odrzucimy jakieś zmienne co jest założeniem dość mocnym. Z drugiej strony, nie jest łatwo posługiwać się zaproponowanym kryterium, gdyż iloraz jest bardzo nieodporny na wartości mianownika bliskie zeru, które mogą powodować fałszywe podziały zbioru wszystkich zmiennych. Alternatywą dla metody wskaźnika $RR(s)$ mogłoby być grupowanie wartości wskaźnika $TOPRI(u)$ metodą k-średnich (dla k=2) gdzie punktami startowymi byłyby dwie wartości skrajne wskaźnika $TOPRI(u)$. Stosując te metodę unikamy problemu niestabilności ilorazu. Dla pojedynczych, w szczególności empirycznych zbiorów danych na ogół stosuje się badanie wizualne wykresu osypiska.

Należy zaznaczyć, że metodę HINoV można modyfikować na wiele sposobów. Na przykład, w początkowym etapie, dla każdej zmiennej v przeprowadzamy wielokrotne grupowanie metodą k-średnich i zapamiętujemy to grupowanie, które miało najmniejszą wariancją wewnątrzgrupową tj.

$$SSE(v) = \sum_{k=1}^{K} \sum_{i \in C_k} \left(x_{iv} - \bar{x}_{kv}\right)^2 \qquad \text{gdzie} \qquad \bar{x}_{iv} = \frac{1}{n_k} \sum_{i \in C_k} x_{iv} \,. \tag{3.30}$$

Tak postąpili we wspomnianym badaniu symulacyjnym Steinley i Brusco, grupując obiekty 50 razy. Metoda HINoV spisała się bardzo dobrze okazując się jedną z trzech najlepszych metod (pamięć - 93% , precyzja - 95% , ARI – 85%).

Metoda HINoV została również zmodyfikowana (*Walesiak i Dudek,* 2008) tak, by można było ją stosować do danych mierzonych na słabszych skalach. Modyfikacja polega na tym, by zamiast grupowania metodą *k*-średnich zastosować metodę aglomeracyjną. Okazało się, że metoda HINoV spisała się dobrze również przypadku porządkowej skali pomiarowej, dla zastosowanych metod grupowania danych osiągając wysokie wartości indeksu Randa dla zbioru zmiennych wybranych w porównaniu z wartościami indeksu Randa dla zbioru wszystkich zmiennych. Modyfikację metody HINoV zaproponowaną przez Walesiaka i Dudka można rozszerzyć na dowolny zbiór zmiennych mierzonych na różnych skalach. Należy w tym celu ujednolicić pomiar odległości pomiędzy obiektami oraz zastosować metodę grupowania, która będzie nadawała się do wszystkich skal pomiarowych.

Algorytm *HINoV* na ogół poprawnie identyfikuje zmienne zakłócające, pozwalając odkryć skupienia w zbiorze danych, jednak ma też pewne ograniczenia. Procedura zależy od relacji pomiędzy pojedynczymi zmiennymi. Jeśli tylko jedna zmienna będzie posiadała zdolność do dyskryminacji analizowanych obiektów, czyli tworzyła strukturę skupień, to algorytm *HINoV* nie wykryje zmiennych zakłócających (por. *Walesiak i Dudek,* 2008). Podobnie w przypadku, jeżeli wszystkie zmienne są nieistotne dla struktury skupień, czyli gdy jej nie ma. Dla tej metody (i dla obu jej modyfikacji por. § 3.7 i § 3.8) dość istotne jest poprawne określenie liczby skupień. Możliwe są dwa rozwiązania: wcześniejsze ustalenie liczby klas co jest raczej niemożliwe przed ustaleniem zbioru zmiennych opisujących obiekty (często błędne jest nawet po ustaleniu zmiennych) lub dokonanie analizy w szerszym ujęciu, ze sprawdzeniem występowania różnych uzasadnionych wariantów liczby skupień, np. w przedziale od 2 do 10 (liczby, które w zastosowaniach praktycznych pojawiają się najczęściej). Można ocenić uzyskane wyniki przy pomocy metod badających stabilność podziału lub analizy replikacji (por. § 1.2) i rozstrzygnąć o faktycznej liczbie skupień przyjmując ostateczne rozwiązanie. Przeprowadzone zostało badanie mające na celu sprawdzenie jakości takiego rozstrzygnięcia. Na zestawie zbiorów PROB zbadano jakość analizy replikacji i indeksu stabilności podziału (1.14) (por. 1.4.1). Wyniki badania pokazały, że indeks oceniający stabilność nie nadaje się do ustalania liczby skupień – indeks ten ocenił prawidłowo liczbę skupień tylko w, około 30% zbiorów. Nieco lepsze rezultaty dała analiza replikacji prawidłowo wskazując liczbę skupień w, około 53% zbiorów. Te wyniki pozwalają zaprojektować badanie efektywności HINoV na zestawie PODST 9450 (por. 1.4.1) zbiorów z użyciem replikacji opartej na grupowaniu metodą *k*-średnich (liczba powtórzeń replikacji równa 20). Badanie efektywności zostało przeprowadzone w niniejszej pracy w celu rzetelnego porównania tej metody z własnymi propozycjami (por. rozdział 4) na tych samych zbiorach danych Wyniki badania zawarte są w tabelach 3.1 oraz 3.2.

Wyniki są nieco inne niż w pracy Steinley i Brusco (2008) ale relatywna ocena HINoV w porównaniu z jej dwiema modyfikacjami (por. § 3.7 i 3.8) jest taka sama. Jakość HINoV i jej dwóch modyfikacji należy ocenić jako podobną. Biorąc pod uwagę to, że HINoV jest bardziej uniwersalną metodą, można by ocenić ją nawet wyżej od obu jej modyfikacji. Niestety, poważną wadą

Tabela 3.1 Efektywność metody HINoV względem liczby zmiennych nieistotnych oraz rodzaju struktury skupień tworzonej przez zmienne istotne.

Typ zbioru danych	Liczba zmiennych nieistotnych	Pamięć	Precyzja	*ARI*
Brak korelacji wewnątrz klas	2	0,804	0,918	0,695
	4	0,741	0,795	0,628
	6	0,687	0,700	0,570
Korelacja wewnątrz klas	2	0,828	0,965	0,645
	4	0,786	0,856	0,599
	6	0,747	0,778	0,561
Średnio		**0,766**	**0,835**	**0,616**

Źródło: obliczenia własne.

Tabela 3.2 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.

tej metody jest zupełny brak odporności na skorelowanie zmiennych nieistotnych dla struktury skupień, co jest widoczne w tabeli 3.2. Badanie

odporności metody HINoV na błędnie zadaną liczbę skupień można też znaleźć w pracy Korzeniewski (2010).

3.7. Metoda VS-KM

Metoda VS-KM (*Variable Selection for K-Means*) (*Brusco i Cradit,* 2001) jest modyfikacją metody HINoV ukierunkowaną na uwzględnienie ewentualnego skorelowania zmiennych nieistotnych. Metoda ta po uwzględnieniu niewielkich modyfikacji wprowadzonej przez Steinley'a i Brusco (*Steinley i Brusco,* 2008) polega na wykonaniu następujących kroków.

Krok 1. Ustalmy *próg*=0 $g_{\min} = 0{,}05$ $g_{fac} = 0{,}5$.

Krok 2. Dla wszystkich zbiorów zmiennych jedno oraz dwuelementowych czyli dla $P \in PP(1) \cup PP(2)$ przeprowadzamy 50 razy grupowanie metodą *k*-średnich i przez $\pi(P)$ oznaczmy podział zbioru danych z najmniejszą wariancją wewnątrzgrupową tj.

$$SSE(\pi(P)) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{v \in P} (x_{iv} - \bar{x}_{iv})^2 \qquad (3.31)$$

Krok 3. Dla wszystkich par zmiennych (u,v), z których każdej odpowiada jakiś podział uzyskany w punkcie 2 znajdujemy $ARI(u,v)$ (por. wzór (3.27)) pomiędzy tymi podziałami. Dla wszystkich podziałów z punktu 2 uzyskanych w oparciu o dwie zmienne (u,v) znajdujemy wskaźnik

$$VAF(\pi(P)) = \frac{TSS(P) - SSE(\pi(P))}{TSS(P)} \qquad (3.32)$$

gdzie $TSS(P) = (n-1)\sigma^2(P)$.

Krok 4. Wybieramy parę zmiennych (u^*, v^*) spełniających kryterium

$$VAF^* = \max_{(u,v)} \{VAF(u,v) : ARI(u,v) > próg\} \qquad (3.33)$$

Ustalmy $s = 2$, $\eta = ARI(u^*, v^*)$, $PP^*(s) = (u^*, v^*)$ i zapamiętujemy $\pi(PP^*(s))$.

Krok 5. Jeśli $s = 2$, to przechodzimy do kroku 6. W przeciwnym przypadku, kładziemy $P = PP^*(s)$ i przeprowadzamy 50 grupowań zbioru

metodą *k*-średnich i oznaczamy przez $\pi\left(PP^*(s)\right)$ ten z otrzymanych podziałów, który ma najmniejszą wariancję wewnątrzgrupową.

Krok 6. Dla każdej zmiennej $v \in PP(V) \setminus PP(s)$ znajdujemy ARI pomiędzy podziałem znalezionym dla zmiennej v w kroku drugim a podziałem $\pi\left(PP^*(s)\right)$. Oznaczamy tę wartość przez $ARI\left(PP^*(s), v\right)$.

Krok 7. Połóżmy $\lambda = \max_v ARI\left(PP^*(s), v\right)$. Jeżeli $\lambda < g_{\min}$ lub $\lambda < \eta \cdot g_{fac}$, to uznajemy zbiór zmiennych $PP^*(s)$ jako zbiór zmiennych istotnych i kończymy postępowanie. W przeciwnym przypadku idziemy do kroku 8.

Krok 8. Połóżmy $\eta = \lambda$, $PP^*(s+1) = PP^*(s) \cup v'$ gdzie v' to zmienna z kroku 7 dająca maksymalną wartość λ, oraz $s = s+1$. Jeśli $s = V$ to uznajemy zbiór zmiennych $PP^*(s)$ jako zbiór zmiennych istotnych i kończymy postępowanie. W przeciwnym przypadku idziemy do kroku 5.

Modyfikacje w stosunku do postaci oryginalnej metody polegają na tym, że : a) wykonujemy (w punkcie 2) wielokrotne grupowania metodą *k*-średnich zamiast zastosowania metody grupowania hierarchicznego Warda do wyznaczenia punktów startowych dla metody *k*-średnich, b) stosujemy *próg*=0 zamiast *próg*=0,25.

Mogłoby się wydawać, że algorytm powyższy powinien mieć przewagę nad metodą HINoV ponieważ w kroku 3 wybieramy podział oparty na dwóch zmiennych o najmniejszej wariancji zewnątrzklasowej w stosunku do wariancji całkowitej (pod warunkiem minimalnej zgodności grupowań opartych na każdej z obu zmiennych) a następnie dołączamy po jednej zmiennej najbardziej zgodnej z dotychczas wybranymi (pod warunkiem, że ubytek *VAF* nie jest zbyt duży). Strategia dołączania pojedynczych zmiennych okazała się jednak decydująca i nie dała żadnej przewagi nad HINoV, gdyż nowy algorytm bardzo dobrze wybierał zmienne istotne dla struktury skupień (pamięć - 98%), ale wybierał ich za dużo bo precyzja - 84% jest wyraźnie gorsza od precyzji metody HINoV (ponadto *ARI* – 85% jest na identycznym poziomie). Przyczyną są zapewne niedoskonałe wielkości kryteriów stopu z kroku 7 odnoszące się do poprawiania wartości wskaźnik *ARI*. Z kolei, gdyby zastosować oryginalnie proponowany *próg*=0,25, to metoda wybierałaby za mało zmiennych, bo w czwartym kroku niekiedy nie wybierałaby żadnych zmiennych (por. *Steinley i Brusco,* 2008). W oryginalnym opracowaniu autorów metody były badane również inne wartości *progu.* Wówczas doszli oni do wniosku, że *próg*=0,25 jest właściwy. Ogólnie rzecz biorąc, wnioskując na podstawie eksperymentu symulacyjnego

przeprowadzonego w cytowanej pracy, trudno uznać tę modyfikację za lepszą od HINoV.

Nieco inne wyniki daje badanie przeprowadzone w niniejszej pracy na zestawie PODST zbiorów przy takich samych parametrach replikacji jak dla HINoV. Szczegółowe wyniki badania zawarte są w tabelach 3.3 i 3.4. Ocena względna metody VS-KM w porównaniu z innymi metodami jest inna od tej z pracy Steinley i Brusco, (2008). Metodę tę należy uznać za lepszą od HINoV – autorom udało się poprawić HINoV pod względem jej wrażliwości na skorelowanie zmiennych nieistotnych, choć nadal efektywność dla takich

Tabela 3.3 Efektywność metody VS-KM względem liczby zmiennych nieistotnych oraz rodzaju struktury skupień tworzonej przez zmienne istotne.

Typ zbioru danych	Liczba zmiennych nieistotnych	Pamięć	Precyzja	*ARI*
Brak korelacji wewnątrz klas	2	0,861	0,928	0,708
	4	0,844	0,899	0,686
	6	0,839	0,876	0,680
Korelacja wewnątrz klas	2	0,883	0,958	0,651
	4	0,876	0,942	0,647
	6	0,870	0,928	0,638
Średnio		**0,862**	**0,922**	**0,668**

Źródło: obliczenia własne.

Tabela 3.4 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.

rozkładów zmiennych zakłócających jest wyraźnie niższa od innych. Nieco gorzej natomiast dla tej modyfikacji wypada ocena oparta na empirycznych zbiorach danych (por. rozdział 5).

3.8. Metoda VAF z ważeniem zmiennych indeksem skupialności

Te metodę zaproponowali Steinley i Brusco (2007(a)). Polega ona na tym, że zanim zastosujemy metodę podobną do VS-KM, to do każdej zmiennej należy dołączyć wagę wynikającą ze wskaźnika skupialności :

$$CI_v = \frac{12 \times \sigma^2(v)}{r^2(v)} \tag{3.34}$$

zaś odpowiadającym mu względnym wskaźnikiem skupialności wskaźnik postaci

$$RC_v = \frac{CI_v}{\min(CI_v)} \quad \text{dla} \quad v = 1,\ldots,V \tag{3.35}$$

Wskaźnik ten, jak widać, odnosi wariancję pojedynczej zmiennej do kwadratu rozstępu tej zmiennej. Taki wskaźnik sam w sobie niewiele da, ale połączony z inną, bardziej zaawansowaną metodą może ją nieco poprawić. Autorzy połączyli ten wskaźnik z rozszerzoną postacią metody VS-KM. Opracowali metodę, którą można scharakteryzować następująco.

Na początku obliczamy wskaźniki dane wzorem (3.34) dla każdej zmiennej. Następnie obserwacje na *j*-ej zmiennej standaryzujemy przy pomocy formuły

$$x_{ij}^* = \frac{x_{ij} - \overline{x}_j}{\sigma_j} \quad \text{dla} \quad i = 1,\ldots,n \quad j = 1,\ldots,V \tag{3.36}$$

Trzecim i ostatnim krokiem transformacji oryginalnego zbioru zmiennych jest zastosowanie przekształcenia

$$t_{ij} = x_{ij}^* \sqrt{\frac{RC_j \cdot \left(r_{j,\min}^*\right)^2}{\left(r_j^*\right)^2}} \quad \text{dla} \quad i = 1,\ldots,n \quad j = 1,\ldots,V \tag{3.37}$$

gdzie $r_{j,\min}^*$ odpowiada zmiennej o najmniejszym wskaźniku (3.35) Takie przekształcenie zmiennych sprowadza wszystkie zmienne do podobnego zakresu zmienności zachowując w pewnym stopniu cechy skupialności poszczególnych zmiennych. Następnym etapem jest wstępna eliminacja zmiennych, która polega na odrzuceniu zmiennych, dla których wskaźnik skupialności (3.34) jest niższy od progu danego liczbą $CI_{0,95 \cdot B}$. Wartość tego progu obliczamy generując ciągów n-elementowych ze standaryzowanego rozkładu normalnego.

W dalszej kolejności, dla dowolnego niepustego podzbioru P zmiennych , o mocy s przeprowadzamy 20 razy grupowanie zbioru danych metodą k-średnich i zapamiętujemy podział zbioru danych z najmniejszą wariancją wewnątrzgrupową $SSE(\pi(P))$. Dla tego podziału znajdujemy

$$VAF(\pi(P)) = \frac{TSS(P) - SSE(\pi(P))}{TSS(P)} \qquad (3.38)$$

Dla każdego $1 \le s \le V$ znajdujemy $VAF^*(s)$ tj. największą spośród wszystkich wartości $VAF(s)$. Mając ustalone $VAF^*(s)$ i, co za tym idzie zbiór zmiennych $P^*(s)$ tę wielkość wyznaczający, spośród s takich, że $1 < s < V$, wybieramy liczbę zmiennych istotnych, podobnie jak w metodzie HINoV tj. kierując się jak największą wartością ilorazu

$$RR(s) = \frac{VAF^*(s) - VAF^*(s+1)}{VAF^*(s-1) - VAF^*(s)} \qquad (3.39)$$

po uprzednim malejącym uporządkowaniu wartości $VAF^*(s)$. Dla niedużej liczby zmiennych możemy stosować tę metodę w wersji przeszukującej wszystkie podzbiory zmiennych zaś dla większej liczby zmiennych autorzy proponują osobny algorytm. Algorytm polega na tym, by zbadać wszystkie możliwe podzbiory zmiennych o liczebnościach nieprzekraczających 3, a następnie dołączać pojedyncze zmienne pod warunkiem, że po dołączeniu otrzymamy większą wartość wyrażenia (3.38).

W badaniu symulacyjnym przeprowadzonym przez Steinley'a i Brusco (2008) metoda VAF wspomagana ważeniem zmiennych indeksem skupialności spisała się na równi z metodami HINoV oraz VS-KM (pamięć - 89% , precyzja - 96% , ARI – 86%). Metoda VAF została zbadana w niniejszej pracy na zestawie PODST zbiorów. Należy zauważyć, że, w przypadku tej metody, wykonanie analizy replikacji w celu ustalenia optymalnej liczby skupień jest bardzo czasochłonne pomimo wykonywania 20 grupowań zamiast 50 (tak jak dla HINoV i VS-KM). Otrzymane wyniki nie potwierdzają tych z badania autorów metody - wskaźniki efektywności są znacznie niższe nie tylko od metody VS-KM ale nawet od HINoV. Szczegółowe wyniki badania zawarte są w tabelach 3.5, 3.6. W uzupełnieniu do tych wyników należy dodać, że we wstępnym etapie selekcji zmiennych wyznaczona została wartość progu $CI_{0,95\cdot B}$ = **0,3087** na podstawie 100 zbiorów składających się z 200 wartości wygenerowanych ze standaryzowanego rozkładu normalnego. Metoda odrzucała zmienne mające wartość indeksu (3.34) poniżej progu, w przypadku około 8% pojedynczych zmiennych. W 15% odrzuceń było to odrzucenie błędne, gdyż odrzucona zmienna była istotna dla struktury skupień. Ten wynik należy ocenić jako słaby.

Należy zauważyć, że trudno poprawić wyznaczoną wartość progu wstępnego etapu selekcji. Gdyby przyjąć wartość niższą to błędnych odrzuceń byłoby więcej. Jeśli zaś wartość progu miałaby być wyższa, to mogłaby ona zmniejszyć

Tabela 3.5 Efektywność metody VAF względem liczby zmiennych nieistotnych oraz rodzaju struktury skupień tworzonej przez zmienne istotne.

Typ zbioru danych	Liczba zmiennych nieistotnych	Pamięć	Precyzja	*ARI*
Brak korelacji wewnątrz klas	2	0,798	0,853	0,656
	4	0,751	0,758	0,617
	6	0,712	0,733	0,600
Korelacja wewnątrz klas	2	0,790	0,883	0,595
	4	0,757	0,808	0,571
	6	0,721	0,791	0,546
Średnio		**0,755**	**0,804**	**0,598**

Źródło: obliczenia własne.

Tabela 3.6 Efektywność metody VAF 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	0,779	0,799	0,801	0,716	0,604	0,790	0,861
Precyzja	0,839	0,836	0,839	0,760	0,640	0,89	0,905
ARI	0,626	0,633	0,634	0,576	0,538	0,622	0,569

Źródło: obliczenia własne.

możliwości dokonywania wstępnej selekcji zmiennych. Konieczność zbadania wszystkich podzbiorów zmiennych zupełnie dyskwalifikuje tę metodę przy dużej liczbie zmiennych. Dla niewielkich liczb zmiennych (tak jak w zestawie PODST) wydłuża to czas obliczeń, a wyniki w stosunku do VS-KM i HINoV nie zostają poprawione – są gorsze. Do podobnych wniosków doszli Fraiman i

inni (2008) porównując tę metodę z VS-KM, mimo iż badanie dotyczyło bardzo prostej i dość wyraźnej struktury 3 skupień. Ponadto, w zastosowaniach empirycznych (por. rozdział 5) metoda wypada raczej słabo. Trudno oprzeć się wrażeniu, że indeks skupialności pojedynczych zmiennych tyleż pomaga w selekcji zmiennych istotnych dla struktury skupień co pogarsza wyniki eliminując we etapie wstępnym wiele zmiennych tworzących strukturę skupień. Pomysłodawcy tej modyfikacji chcieli uzyskać możliwość zbadania wszystkich podzbiorów zbioru wszystkich zmiennych kosztem wstępnej eliminacji niektórych pojedynczych zmiennych oraz kosztem zmniejszenia jakości grupowania. Ten pomysł zupełnie się nie sprawdził nawet dla zbiorów z niewielką liczbą zmiennych. Tę modyfikację metody HINoV należy ocenić negatywnie.

3.9. Metody oparte na entropii

Entropię zbioru obserwacji $x_1,...,x_n$ definiujemy jako (por. *Dash i Liu* , 2000)

$$E(x_1,...,x_n) = \sum_{x_1 x_2 ... x_n} p(x_1,...,x_n)\log(p(x_1,...,x_n)) \qquad (3.40)$$

Im większa entropia tym rozkład obserwacji bardziej równomierny, im wyraźniejsza struktura skupień w zbiorze danych tym mniejsza entropia. Jeżeli istnieje wyraźna struktura skupień, to odległości pomiędzy dwoma punktami są albo duże albo małe – mniej jest odległości pośrednich, choć to zależy od struktury skupień. Załóżmy, że odległości pomiędzy dwoma punktami zostały unormowane oddzielnie na każdej ze zmiennych za pomocą podzielenia odległości przez maksymalną odległość na danej zmiennej (rozstęp zmiennej). Entropię układu dwóch obserwacji znajdujących się w odległości d przybliżamy w następujący sposób

$$E(x_1, x_2) = -d \log d - (1-d)\log(1-d) \qquad (3.41)$$

tak, by wartość maksymalną równą 1 entropia miała dla średniej odległości tzn. d=0,5 , zaś wartość minimalną równą 0 entropia miała dla odległości możliwie najmniejszej tzn. d=0 oraz możliwie największej d=1. Idąc dalej, entropia całego zbioru danych to

$$E = -\sum_{x_1, x_2} [d(x_1, x_2)\log d(x_1, x_2) + (1 - d(x_1, x_2))\log(1 - d(x_1, x_2))] \qquad (3.42)$$

gdzie sumowanie przebiega po wszystkich parach obserwacji ze zbioru danych. Następnie przechodzimy z odległości $d(x_1, x_2)$ pomiędzy obserwacjami na

podobieństwa $S(x_1, x_2)$ obserwacji przyjmujące wartości z przedziału $[0,1]$ przy pomocy wzoru

$$S(x_1, x_2) = \exp(-\alpha \cdot d(x_1, x_2)), \tag{3.43}$$

gdzie α jest tak dobrane, by średnia arytmetyczna ze wszystkich odległości odpowiadała podobieństwu równemu 0,5 czyli $0{,}5 = \exp(-\alpha \cdot \overline{d})$. Zastosowanie podobieństw zamiast odległości prowadzi do tego, że entropia będzie niska, jeżeli podobieństwo będzie niskie lub wysokie, czyli bliskie 0 lub 1. Jeśli podobieństwo jest średnie, czyli w okolicach 0,5 to entropia jest wysoka. A zatem, entropia całego zbioru danych będzie dana wzorem

$$E = -\sum_{x_1, x_2} \left[S(x_1, x_2) \log S(x_1, x_2) + (1 - S(x_1, x_2)) \log(1 - S(x_1, x_2)) \right] \tag{3.44}$$

Najprostszy algorytm wybierania zmiennych oparty na entropii – nazwijmy go pierwszym algorytmem, jaki proponują autorzy, polega na tym, by porównywać entropie wszystkich obserwacji ograniczając się do zbioru wszystkich zmiennych oprócz pojedynczych, kolejnych zmiennych. Jeśli np. po wyrzuceniu pierwszej zmiennej entropia układu wszystkich obserwacji będzie większa od entropii po wyrzuceniu drugiej zmiennej, to oznacza to, że pierwsza zmienna jest bardziej istotna dla struktury danych od drugiej zmiennej. Pozostaje problem ustalenia, w którym miejscu podzielić ciąg uporządkowanych entropii odpowiadających pojedynczym zmiennym. Najprostszym sposobem jest

Tabela 3.7 Efektywność porządkowania zmiennych opartego na entropii.

Typ zbioru danych	Liczba zmiennych nieistotnych	Odsetek prawidłowych uporządkowań	Odsetek uporządkowań z jednym błędem
Brak korelacji wewnątrz klas	2	0,420	0,598
	4	0,350	0,499
	6	0,217	0,348
Korelacja wewnątrz klas	2	0,479	0,640
	4	0,342	0,502
	6	0,260	0,423

Źródło: obliczenia własne.

zastosowanie kryterium ilorazowego (największego skoku) takiego jak np. we wzorze (3.29) lub pogrupowania (propozycja własna) wszystkich wartości entropii metodą *k*-średnich (dla *k*=2) gdzie punktami startowymi byłyby dwie wartości skrajne. Stosując tę metodę unikamy problemu niestabilności ilorazu, jednak pozostaje założenie o tym, że jakieś zmienne musimy wyrzucić (w tym przypadku co najmniej jedną, zaś w przypadku kryterium największego skoku, co najmniej dwie). Taka metoda została zastosowana w pracy Korzeniewski (2009), efektywność metody okazała się słaba.

Innym sposobem wykorzystania entropii do wybierania zmiennych może być algorytm (typu *wrapper* – nazwijmy go drugim algorytmem) jaki proponują autorzy, polegający na zastosowaniu jakiegoś kryterium optymalizacyjnego w połączeniu z jakąś metodą grupowania obserwacji. Autorzy zastosowali grupowanie metodą *k*-średnich (*k* musi być znane) dla wszystkich możliwych podzbiorów zmiennych i wybierali podzbiór, który dawał największą wartość kryterium (1.18). Inną alternatywą (w przypadku dużej liczby zmiennych) może być badanie tylko zbiorów zmiennych uwzględniających uporządkowanie zmiennych wynikające ze zbadania pojedynczych zmiennych (por. pierwszy algorytm). Taką metodę zbadali na kilku (sztucznych i empirycznych) zbiorach danych. Metoda wypadła dobrze ale badanie było bardzo skrótowe, więc nie można wyciągać z niego daleko idących wniosków. Ponadto, metoda ma zasadnicze ograniczenie polegające na tym, że dla zbiorów z większą liczba zmiennych zbadanie wszystkich możliwych podzbiorów zmiennych jest trudne obliczeniowo. Jeśli zaś wykorzystamy uporządkowanie zmiennych wynikające z badania pojedynczych zmiennych, to wraz z konsekwencjami błędnego uporządkowania.

Zbadana została najpierw jakość porządkowania zmiennych w eksperymencie przeprowadzonym na zestawie zbiorów PROB (por. 1.4.1). Jakość oceniana była przy pomocy odsetka prawidłowych uporządkowań wszystkich zmiennych oraz odsetka uporządkowań z jednym błędem, który rozumiemy w ten sposób, że na *k* pierwszych miejscach (*k* – liczba zmiennych istotnych) mogła być co najwyżej jedna zmienna nieistotna dla struktury skupień. *Z* liczb zawartych w tabeli 3.7 wynika, że porządkowanie zmiennych oparte na entropii osiągnęło słabe wyniki. Zauważmy, że z badania tego można wyciągnąć wnioski o jakości drugiego algorytmu w wersji wykorzystującej uporządkowanie istotności zmiennych wynikające z odpowiadających im wielkości entropii. Słaba jakość porządkowania zmiennych spowoduje to, że jakakolwiek metoda wykorzystująca to uporządkowanie w celu ograniczenia liczby koniecznych do zbadania podzbiorów zmiennych, nie da dobrych rezultatów. Przyczyną słabej efektywności metod opartych na entropii jest to, że rozkłady jednomodalne nie tworzące struktury skupień tworzą zbiory danych o

wysokim stopniu uporządkowania, co pociąga za sobą niską entropię, co nie pozwala odróżnić tych zbiorów od zbiorów ze strukturą skupień.

3.10. Metoda uśredniania zmiennych

W roku 2008 zaproponowano (Fraiman i inni, 2008) dość ogólną ideę wybierania zmiennych nieistotnych, którą można zastosować zarówno w analizie skupień jak i w klasyfikacji ze zbiorem uczącym. Punktem wyjścia jest założenie, że zmienne nieistotne dla struktury skupień powinny mieć we wszystkich skupieniach takie same rozkłady i, co za tym idzie, wartości średnie. Wobec tego, mając dany podział zbioru danych 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, z pozostałymi zmiennymi równymi swoim wartościom średnim. Można to zrobić wykorzystując, tak jak w innych metodach, indeksy podobieństwa podziałów ale tu proponowana jest inna metoda. Metodę można opisać dokładniej w następujący sposób.

Niech $X_i = [v_{i1}, \ldots, v_{iV}]^T$ oznacza wektor, którego realizacją jest x_i czyli *i*-ty wiersz macierzy danych (*i*-ta obserwacja). Dla $A \subset VV$ zdefiniujmy wektor X_i^* odpowiadający wektorowi X_i w następujący sposób

$$X_i^* = \begin{bmatrix} v_{i1}^* \\ v_{i2}^* \\ \\ v_{iV}^* \end{bmatrix}, \quad \text{gdzie} \quad v_{ij}^* = \begin{cases} v_{ij} & \text{gdy} \quad j \in A \\ E(v_{ij}) & \text{gdy} \quad j \notin A \end{cases} \tag{3.45}$$

Niech f będzie funkcją przypisującą skupienia wektorowi wszystkich zmiennych ***V***. Zbiór *A* należ wybrać tak, by maksymalizował wartość funkcji

$$h(A) = \sum_{k=1}^{K} P\left(f(X) = k, f(X^*) = k\right) \tag{3.46}$$

gdzie prawdopodobieństwo odnosi się do przestrzeni, na której określone są wektory X_i. Maksymalizując funkcję (3.46) na zbiorze danych otrzymamy regułę mówiącą o tym, że zbiór *A* należy wybierać tak, by maksymalizował wartość funkcji

$$h_n(A) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} \chi\left(f_n(x_i) = k\right) \chi\left(f_n(x_i^*) = k\right) \tag{3.47}$$

gdzie $\chi\left(f_n\left(x_i\right)=k\right)$ oznacza funkcję charakterystyczną *i*-ej obserwacji tj. funkcję przyjmującą wartość 1 gdy obserwacja ta zostanie przypisana do skupienia o numerze *k*, zaś wartość 0 w przeciwnym przypadku. Funkcja f_n (empiryczny odpowiednik funkcji f) jest funkcją przypisującą skupienia obserwacjom odpowiadającą zbiorowi *VV* wszystkich zmiennych tzn. funkcja ta ustala do którego skupienia (przy danym podziale zbioru na skupienia) przypisać obserwację w zależności od wartości wszystkich zmiennych obserwacji. Regułę daną wzorem (3.47) łatwo zastosować gdy metodą grupowania jest metoda *k*-średnich. Wówczas wartość funkcji $\chi\left(f_n\left(x_i^*\right)=k\right)$ jest równa 1 gdy obserwacja x_i^* z uśrednionymi wartościami współrzędnych spoza zbioru *A* ma najbliżej do centrum skupienia o numerze *k*. Autorzy zastosowali tę metodę do ustalenia zmiennej nieistotnej, wybranej z trzech zmiennych, z których dwie pierwsze tworzyły strukturę trzech dobrze separowalnych skupień. W przypadku gdy trzecia zmienna była niezależna od dwóch pierwszych metoda ta spisała się bardzo dobrze okazując się lepszą od dwóch modyfikacji metody HINoV – metody VS-KM i metody VAF z indeksem skupialności. Rozważano również trzecią zmienną w postaci sumy dwóch pierwszych zmiennych i okazało się, że w tym przypadku, metoda nie poradziła sobie z wykryciem trzeciej zmiennej jako nieistotnej. Zauważmy, że przypadek trzeciej zmiennej zależnej w sposób funkcyjny od dwóch pierwszych jest nieco szerszy od przyjętego w tej pracy rozumienia problemu selekcji zmiennych istotnych dla struktury skupień. W tym bowiem przypadku każde dwie zmienne z rozpatrywanej trójki tworzą strukturę skupień i kwestia istotności zmiennych dla struktury skupień jest rozważana łącznie z kwestią eliminowania zmiennych zbędnych (*redundant* – por. § 1.3). Ten problem jest bardziej złożony. Jeżeli zmienna jest zależna funkcyjnie od jakichś innych zmiennych, to niewątpliwie należy ją uznać za zbędną ale nie wiadomo co zrobić jeśli jest bardzo silnie skorelowana z innymi zmiennymi.

Idąc dalej, zaproponowano modyfikację metody danej wzorami (3.45)-(3.47) mając na celu poprawienie efektywności nawet w sensie odrzucenia przy okazji selekcji zmiennych istotnych dla struktury skupień, zmiennych zbędnych. Modyfikacja polega na tym, by definiując uśrednianie zmiennych wzorem (3.47), ograniczyć się do otoczenia lokalnego, na przykład 5 najbliższych sąsiadów obiektu, który chcemy przypisać do jakiegoś skupienia. Taka modyfikacja radzi sobie zarówno z zadaniem selekcji zmiennych istotnych dla struktury skupień jak i z wyeliminowaniem zmiennych zbędnych. Jednak autorzy doszli do takiego wniosku na podstawie eksperymentu, w którym rozważano bardzo wyraźnie separowalne skupienia.

Metoda została zbadana na zestawie PODST zbiorów w pracy Korzeniewski (2012). Okazało się, że osiągnęła słabe wyniki spisując się dobrze tylko w przypadku wyraźnych struktur skupień (*overlap* równy 0 i 0,1).

3.11. Metody oparte na grupowaniu rzadkim

Termin grupowanie rzadkie oznacza grupowanie oparte tylko na niektórych zmiennych spośród wielu opisujących obiekty. Tak więc, w kontekście selekcji zmiennych w analizie skupień, odnosi się właśnie do wybierania zmiennych. Nazwa metody opracowanej przez Witten i Tibshiraniego (2010), zaproponowana w tytule paragrafu, zawiera jednak podtekst warunku oznaczającego, że liczba wszystkich zmiennych jest duża, na ogół dużo wyższa od liczby obiektów tzn. $V >> n$, i z myślą o takich zbiorach danych została skonstruowana, choć, jak twierdzą autorzy, można ją stosować również gdy obiektów jest więcej niż zmiennych.

Ideę metody najłatwiej jest przedstawić wychodząc od standardowej metody *k*-średnich grupowania obiektów, którą można ująć (por. *Steinley,* 2006) w formie warunek minimalizacji wyrażenia

$$\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \sum_{v=1}^{V} \left(x_{iv} - x_{jv}\right)^2 \qquad (3.48)$$

Ten warunek przekształcamy następnie do postaci, w której maksymalizujemy wyrażenie, w którym każda zmienna ma nadaną wagę w_v, czyli należy znaleźć

$$\max_{C_k, w_v} \left\{ \sum_{v=1}^{V} w_v \left(\frac{1}{n} \sum_{i,j} \left(x_{iv} - x_{jv}\right)^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \left(x_{iv} - x_{jv}\right)^2 \right) \right\} \qquad (3.49)$$

przy warunkach brzegowych $w_1^2 + w_2^2 + \ldots + w_V^2 \le 1$, oraz $w_1 + w_2 + \ldots + w_V \le \lambda$, $w_v \ge 0$. Wartość parametru λ (*tuning parameter*) będzie należało dobrać. Wagi zmiennych będą istotnie rzadkie, tzn. będzie można wykorzystać je do selekcji zmiennych, jeśli parametr λ zostanie właściwie dobrany. Warunkiem koniecznym do tego jest

$$1 \le \lambda \le \sqrt{V} \qquad (3.50)$$

Gdy wszystkie wagi zmiennych są równe, to metoda jest równoważna standardowej metodzie *k*-średnich. Ogólnie rzecz biorąc bardzo trudno jest znaleźć optimum globalne, wobec czego autorzy proponują następujący algorytm dwustopniowy oddzielnie optymalizujący wartość wyrażenia (3.49) względem podziału obiektów na skupienia i oddzielnie względem wag w_v.

Krok 1. Położyć wszystkie wagi jednakowe równe $1/\sqrt{V}$.

Krok 2. Wykonywać kolejne iteracje zmieniające wartości (3.51) oraz (3.52) aż do uzyskania zbieżności.

Dla ustalonych wag znaleźć maksymalną wartość wyrażenia (3.49) względem podziału na skupienia czyli znaleźć

$$\min_{C_k}\left\{\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,j\in C_k}\sum_{v=1}^{V}w_v\left(x_{iv}-x_{jv}\right)^2\right\} \tag{3.51}$$

przy pomocy wersji *k*-średnich z ważoną odległością euklidesową.

Dla ustalonego podziału na skupienia znaleźć optymalną wartość wyrażenia (3.49) względem wag w_v za pomocą wzorów

$$w_v=\frac{S(a_v,\Delta)}{\sqrt{\left(S^2(a_1,\Delta)+S^2(a_2,\Delta)+\ldots+S^2(a_V,\Delta)\right)}} \tag{3.52}$$

gdzie

$$a_v=\frac{1}{n}\sum_{i,j}\left(x_{iv}-x_{jv}\right)^2-\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,j\in C_k}\left(x_{iv}-x_{jv}\right)^2 \tag{3.53}$$

oraz $\Delta=0$ jeżeli suma wag będzie mniejsza od λ, zaś w przeciwnym przypadku należy dobrać wartość $\Delta>0$ tak by suma wag była równa λ. Symbol $S(a_v,\Delta)$ oznacza operator miękkiego progu (*soft threshold*), który definiujemy następująco.

$$S(x,\Delta)=sign(x)\left(|x|-\Delta\right)_+ \tag{3.54}$$

gdzie $(t)_+$ oznacza dodatnią część liczby *t*.

Po uzyskaniu zbieżności otrzymujemy podział zbioru na skupienia oraz odpowiadające temu podziałowi wagi w_v. Taki algorytm uwarunkowany jest znajomością parametru λ. Zaproponowany został algorytm doboru wartości tego parametru oparty na statystyce przerwy (*gap statistic*) podobnej do tej, która służy do obliczania indeksu liczby skupień (*Tibshirani i inni,* 2001).

Zauważmy, że problem optymalizacji wyrażenia (3.49) jest bardzo podobny do optymalizacji wyrażenia (3.23) w metodzie COSA, przy czym, w metodzie COSA wagi niektórych zmiennych są bardzo rzadko równe zeru. W metodzie proponowanej przez Witten i Tibshiraniego ten efekt uzyskuje się częściej dzięki lepiej dobranym warunkom brzegowym, w których wagi zmiennych występują zarówno w pierwszej jak i drugiej potędze. Obok scharakteryzowanego powyżej grupowania rzadkiego metodą *k*-średnich opracowano również wariant takiego grupowania dla metod aglomeracyjnych.

Metoda została zbadana w artykule źródłowym pomysłodawców, ale badanie miało zupełnie inny charakter od zazwyczaj stosowanego. Specyfika badania symulacyjnego była podyktowana ukierunkowaniem metody na problem selekcji zmiennych w zastosowaniu do badania zbiorów empirycznych bardzo małej liczby (kilkadziesiąt) przypadków nowotworu opisanych przez bardzo dużą liczbę genów (kilka tysięcy). Rozważano trzy skupienia dane wielowymiarowymi rozkładami normalnymi z jednakowym odchyleniem standardowym na każdej zmiennej, przy czym liczba zmiennych miała wartości od 50 do 1000. Narzucono z góry liczbę zmiennych, które należy wybrać jako tworzące strukturę skupień i sprawdzano przy pomocy odsetka błędów CER (por. § 1.3) jakość metody oraz kilku innych konkurencyjnych. Nowa metoda osiągnęła lepsze wyniki od innych, niektóre odsetki błędów okazały się nawet w istotny statystycznie sposób (test dla dwóch średnich) niższe od odsetków błędów metod konkurencyjnych. Ten specyficzny charakter metody narzuca ograniczone możliwości zastosowania jej do zbiorów danych o charakterze ekonomiczno-socjologicznym.

3.12. Metoda Ichino oparta na teorii grafów

Metodą o bardzo oryginalnym w porównaniu z innymi metodami charakterze jest podejście Ichino, który opracował bardzo ogólny model populacji - model przestrzeni kartezjańskiej (*Cartesian Space Model*, Ichnio, 1985). Model ten może być stosowany do zbiorów danych z szerokim zakresem kategorii zmiennych, od ciągłych do nominalnych. Model oparty jest na pojęciu sumy kartezjańskiej obiektów $\oplus$ oraz iloczynu kartezjańskiego obiektów $\otimes$. Modelem przestrzeni kartezjańskiej nazywamy trójkę $\left(U^d, \oplus, \otimes\right)$, gdzie U^d jest *d*-wymiarową przestrzenią wartości zmiennych losowych opisujących obiekty zbioru danych,

$$U^d = U_1 \times U_2 \times \ldots \times U_d \tag{3.55}$$

natomiast sumę kartezjańską i iloczyn definiujemy w następujący sposób. Sumą kartezjańską dwóch obiektów A oraz B definiujemy w postaci iloczynu kartezjańskiego sum dwóch obiektów na poszczególnych współrzędnych

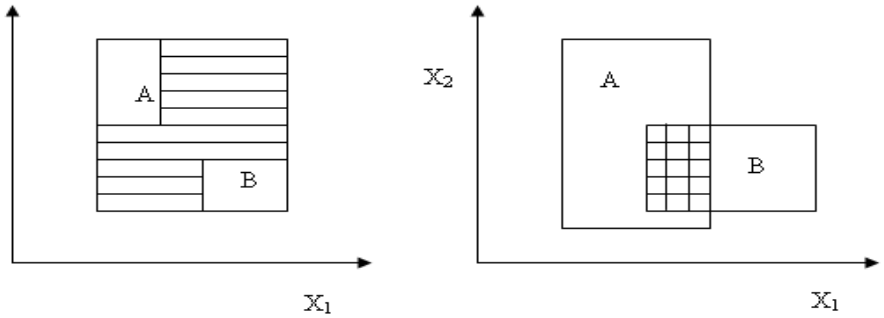
$$A \oplus B = \left(A_1 \oplus B_1\right) \times \left(A_2 \oplus B_2\right) \times \ldots \times \left(A_d \oplus B_d\right) \tag{3.56}$$

Iloczyn kartezjański dwóch obiektów A oraz B definiujemy w postaci iloczynu kartezjańskiego iloczynów dwóch obiektów na poszczególnych współrzędnych

$$A \otimes B = (A_1 \otimes B_1) \times (A_2 \otimes B_2) \times \ldots \times (A_d \otimes B_d) \qquad (3.57)$$

Na przykład, gdy zmienne są mierzone na skali nominalnej, to suma dwóch obiektów na każdej współrzędnej jest sumą mnogościową tych obiektów, gdy zaś obiekty są przedziałami liczbowymi, to suma dwóch obiektów jest przedziałem, którego lewy kraniec jest mniejszym z dwóch lewych krańców a prawy kraniec jest większym z dwóch prawych krańców. Iloczyn dwóch obiektów na każdej współrzędnej jest iloczynem mnogościowym.

Na rysunku 3.1 przedstawiona jest przykładowa ilustracja sumy kartezjańskiej dwóch obiektów i iloczynu kartezjańskiego dwóch obiektów opisanych przez dwie zmienne będące przedziałami liczbowymi. W dalszym ciągu definiujemy obiekty wzajemnie sąsiadujące (*mutual neighbours*) A_1 oraz A_2 względem zbioru obiektów symbolicznych $B = \{B_1, B_2, \ldots, B_m\}$ jeśli spełniony jest warunek



Rys. 3.1 Graficzna ilustracja sumy kartezjańskiej dwóch obiektów (obszar zakreskowany wraz z dwoma obiektami na lewym rysunku) i iloczynu kartezjańskiego dwóch obiektów (obszar kratkowany na prawym rysunku).

Źródło: Ichino (1985).

$$\forall B_i \in B, \quad B_i \otimes (A_1 \oplus A_2) = \varnothing \qquad (3.58)$$

Zbiór obiektów symbolicznych $A = \{A_1, A_2, \ldots, A_k\}$ nazywamy grafem wzajemnego sąsiedztwa (*mutual neighbourhood graph*) względem zbioru

obiektów symbolicznych $B = \{B_1, B_2, \ldots, B_m\}$, gdy każda para obiektów ze zbioru *A* jest wzajemnie sąsiadująca względem zbioru *B*. Grafy wzajemnego sąsiedztwa pozwalają opisywać separowalne skupienia obiektów.

Ichino bardzo dobrze pracował aparat matematyczny, który jest podstawą jego metody udowadniając wiele własności grafów wzajemnego sąsiedztwa. Niestety, o wiele słabiej przedstawione są możliwości zastosowań jego metody. W literaturze znaleźć można tylko dwa przykłady zastosowania (autor nie podaje nawet wszystkich szczegółów) do zbiorów danych empirycznych ze zmiennymi mierzonymi na skali nominalnej.

Pełka i Wilk (2010) sformułowali algorytm metody Ichino w następujący sposób.

Krok 1. Dla każdej kombinacji *P* zmiennych spośród wszystkich $\binom{V}{P}$ kombinacji znajdujemy wszystkie grafy wzajemnego sąsiedztwa .

Krok 2. Znajdujemy liczbę wszystkich par obiektów wewnątrz grafów (*m* – liczba obiektów w grafie, $\frac{1}{2}m(m-1)$ - liczba wszystkich par obiektów).

Krok 3. Dla każdej liczby zmiennych od 1 do *V* wybieramy tę kombinację zmiennych, dla której suma liczb znalezionych w punkcie 2 jest największa.

Krok 4. Wybieramy jako zmienne istotne dla struktury skupień te *k* zmiennych, które wchodzą w skład kombinacji dającej największy przyrost liczby z punktu 2 w stosunku do *k*-1 zmiennych.

Następnie, porównali metodę Ichino rozumianą w sensie powyżej opisanego algorytmu z metodą HINoV (por. § 3.6) na przykładzie 4 zbiorów danych ciągłych zawierających struktury kilku skupień o różnym stopniu separowalności, kształcie, opisanych przez kilka zmiennych wraz z dodanymi dwiema zmiennymi nieistotnymi dla struktur skupień. Autorzy ocenili metodę Ichnio jako nieefektywną (w porównaniu z HINoV) z powodu silnego ograniczenia natury obliczeniowej, które skutkuje niemożnością rozważenia większej liczby zmiennych. Nie wiadomo jednak czy taka ocena jest obiektywna, bo HINoV można, co prawda, zastosować do bardzo dużej liczby zmiennych, ale metoda ta spisuje się wtedy bardzo słabo.

Metoda Ichino została również zbadana w pracy Korzeniewski (2012). To badanie miało nieco inny charakter. Celem było ocenienie efektywności metody w zastosowaniu do zbiorów danych zawierających tylko zmienne ciągłe. Ponadto, metoda została sformułowana nieco inaczej niż w pracy Pełki i Wilk, będąc bardziej podobną do dwóch przykładów podanych przez Ichino poprzez

początkowe pogrupowanie obiektów w ustaloną liczbę skupień. Dokładne ujęcie metody było następujące.

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

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

Krok 3. Za najlepiej odzwierciedlający strukturę skupień uznajemy ten podzbiór *s* zmiennych, dla którego różnica $l(s)-l(s-1)$ ma wartość najwyższą.

Różnice w stosunku do algorytmu z pracy Pełki i Wilk są dość istotne, ale wnioski były podobne – tak ujęta metoda okazała się bardzo nieefektywna dla zbiorów danych ciągłych.

Najważniejszą zaletą metody Ichino opartej na teorii grafów jest jej ogólność w sensie możliwości zastosowania do zbiorów danych ze zmiennymi różnych rodzajów, jednak odnosi się wrażenie, że autor skonstruował ją raczej z myślą o słabych skalach pomiarowych. Bardzo poważnym ograniczeniem metody jest konieczność zbadania wszystkich podzbiorów zbioru wszystkich zmiennych.

4. NOWA METOD SELEKCJI ZMIENNYCH W ANALIZIE SKUPIEŃ

4.1. Korelacja odległościowa i jej własności

Wybieranie zmiennych tworzących strukturę skupień wiąże się ze znalezieniem odpowiedzi na pytanie o to czy dla pewnych zmiennych można pogrupować obiekty zbioru w skupienia czyli spójne podzbiory obiektów o większej gęstości lokalnej od gęstości innych podzbiorów zbioru danych. Jak wynika z przeglądu istniejących metod z rozdziałów 2 i 3, najczęściej istnienie struktury skupień badamy przez próby pogrupowania obiektów w skupienia. Takie metody mają ograniczoną stosowalność a ponadto bardzo trudno połączyć efektywne grupowanie obiektów ze znalezieniem kryterium, które posiadałoby własność monotoniczności względem rodzaju zmiennych dołączanych (lub zabieranych) z rozważanego na danym etapie metody zbioru zmiennych. Zauważmy, że ogólnie rzecz biorąc, jeśli jakieś dwa rozłączne zbiory zmiennych są istotne dla struktury skupień, to włączenie pewnych podzbiorów obiektów do obszaru o większej gęstości, czyli do skupienia, powinno odbywać się „na obu" zbiorach zmiennych. Innymi słowy, jeśli w zbiorze danych istnieje wyraźna struktura skupień, to odległości pomiędzy parami obserwacji obliczane w oparciu o te dwa zbiory zmiennych powinny być ze sobą skorelowane dodatnio. Dla pary obiektów z tego samego skupienia bowiem, odległość obliczona w oparciu o jakikolwiek zbiór zmiennych istotnych dla struktury skupień powinna być mała zaś dla pary obserwacji z różnych skupień odległość powinna być duża. Może się oczywiście zdarzyć, że, lokalnie, dla pewnych skupień ta zasada będzie naruszona (np. przy małych liczebnościach zbiorów zmiennych, lub gdy niektóre skupienia są dużo większe w sensie odległości i liczebności od innych) ale, dla wszystkich par skupień naruszenie tej zasady jest bardzo mało prawdopodobne. Sformułujmy zatem definicję współczynnika korelacji, który będzie pomocny w wykrywaniu takich zależności.

Definicja 4.1 Współczynnik korelacji odległościowej pomiędzy zbiorami zmiennych $A, B \subset VV$ dany jest wzorem

$$WKO(A,B,l) = \frac{\frac{1}{l}\sum_{i=1}^{l} d_i^A d_i^B - \overline{d}^A \overline{d}^B}{s^A s^B}, \qquad (4.1)$$

gdzie $1 \le l \le n$ oznacza liczbę par obserwacji wylosowaną w sposób zależny spośród wszystkich par obserwacji; d_i^A, d_i^B oznaczają odległości dla *i*-tej pary

obliczone w oparciu o zmienne ze zbioru, odpowiednio, *A, B*; $\overline{d}^A$, $\overline{d}^B$, s^A, s^B oznaczają, odpowiednio, średnie arytmetyczne i odchylenia standardowe obliczone spośród wszystkich *l* odległości.

Uwaga 1. Współczynnik korelacji odległościowej (*WKO*) zależy od liczby *l* wylosowanych par, ale liczba ta będzie , na ogół ustalona jako $l = 30$ jeśli nie będzie zaznaczone inaczej. Wówczas, w skrócie, zamiast $WKO(A, B, l)$ będziemy pisać $WKO(A, B)$.

Uwaga 2. *WKO* ma postać współczynnika korelacji liniowej pomiędzy odległościami obliczonymi w oparciu o różne zbiory zmiennych oraz na podstawie wylosowanej próby obserwacji. Celem wyrażenia podanego w formule (4.1) nie jest jednak szacowanie odnośnego współczynnika korelacji odległościowej (wówczas liczba wylosowanych par obserwacji musiałaby być dużo wyższa niż 30). Współczynnik ten wykorzystamy jedynie do wykrycia zależności pomiędzy zbiorami zmiennych dla których jest on obliczany a ewentualną strukturą skupień w zbiorze danych.

Uwaga 3. *WKO* szacuje siłę korelacji liniowej pomiędzy odległościami obliczonymi w oparciu o różne zbiory zmiennych. Jeśli korelacja taka faktycznie istnieje i jest dodatnia, to, oczywiście, nie musi mieć postaci czysto liniowej. Jednak jeśli zależność ma charakter krzywoliniowy, to *WKO* dany wzorem (4.1) powinien być również w tym przypadku dodatni.

Uwaga 4. Współczynnik korelacji odległościowej jest przemienny w stosunku do zbiorów zmiennych tzn. $WKO(A, B) = WKO(B, A)$.

Uwaga 5. W definicji *WKO* nie precyzujemy tego czy zbiory A, B są rozłączne czy nie. Należy jednak pamiętać o tym, że jeżeli nie są rozłączne, to zdolność *WKO* do wykrywania zależności pomiędzy zbiorami zmiennych a strukturą skupień maleje gdyż występuje powielanie informacji.

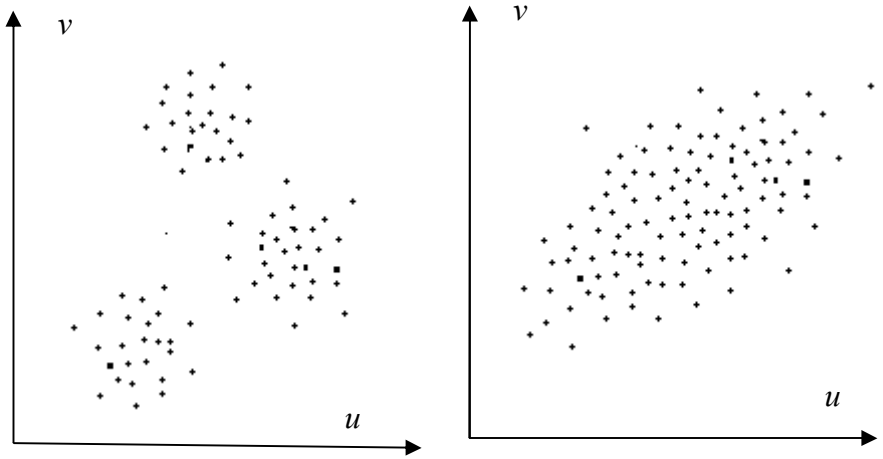
Uwaga 6. Jednokrotne obliczenie wartości *WKO* nawet dla większej niż 30 liczby par nie wystarcza, ponieważ wartości te są bardzo niestabilne, są silnie uzależnione od liczby wylosowanych par obserwacji. W celu wyznaczenia stabilnej średniej wartości *WKO* obliczanie jej dla 30 par należy wielokrotnie powtórzyć. Możemy zmodyfikować proces losowego doboru par tak, by wylosowane pary stanowiły jak najbardziej reprezentatywną próbę ze zbioru wszystkich par oraz w celu skrócenia czasu obliczeń. Załóżmy, że w zależności od tego jak liczny będzie zbiór obiektów kandydatów do wylosowania (na ogół cały zbiór danych), to stosować będziemy trzy różne sposoby obliczania wartości *WKO*. Pierwszy sposób to wylosowanie 30 obiektów do każdego z *n* obiektów zbioru. Otrzymujemy wówczas *n* podzbiorów składających się z 30 par obiektów (obiekty w każdej z 30 par są różne jednym z elementów pary jest

ten sam obiekt). Drugi sposób (stosowany przy dużej liczbie obiektów w zbiorze) to wylosowanie 300 podzbiorów 30 obiektów (również tak by w każdej z 30 par obiekty były różne). Trzeci sposób to wariant kompletny, stosowany przy małej liczbie obiektów w zbiorze. Polega on na obliczeniu wartości współczynnika na podstawie wszystkich $0{,}5 \cdot n(n-1)$ par obiektów.

Jak wykorzystać zaproponowaną definicję *WKO* do precyzyjnego ustalenia zależności pomiędzy zbiorami zmiennych a ewentualną strukturą skupień ? Ustalenie analityczne rozkładu prawdopodobieństwa współczynnika korelacji odległościowej nie jest łatwe. O współczynniku korelacji linowej wiadomo tylko tyle, że jeżeli zmienne mają rozkłady normalne, to współczynnik korelacji liniowej ma rozkład *t*-Studenta (w dużej próbie normalny). Niestety, niewiele wiadomo o rozkładzie odległości pomiędzy parami obiektów. Jeżeli rozkłady wszystkich zmiennych są jednomodalne lub równomierne, to rozkład odległości pomiędzy parami obiektów jest jednomodalny, co można ustalić symulacyjnie (por. *Dash i Liu,* 2000). Jeżeli zmienne tworzą strukturę skupień, to często zdarza się, że rozkład odległości pomiędzy parami obiektów jest wielomodalny, na ogół, silnie asymetryczny, ale trudno uzyskać jakieś konkretne wnioski badawcze. Dla zmiennych jednomodalnych nieposiadających struktury skupień rozkład odległości pomiędzy parami obiektów również może być silnie asymetryczny. Ponadto, przy większej liczbie skupień, rozkład odległości pomiędzy parami obiektów traci silną asymetrię i zaczyna przypominać rozkład odległości dla obiektów o rozkładzie równomiernym.

Dodatnia wartość *WKO* pomiędzy dwiema zmiennymi może wynikać z istotności tych zmiennych dla struktury skupień w zbiorze danych ale może też być wynikiem skorelowania zmiennych, które mogą być nieistotne dla struktury skupień (por. rys. 4.1). Jak rozróżnić te przypadki?

Sformułujemy kilka podstawowych własności współczynnika korelacji odległościowej w oparciu o przykłady struktur skupień w postaci mieszanek rozkładów normalnych. Prześledźmy zachowanie się *WKO*(*u,v*) dla zmiennych *u, v* ciągłych, w zależności od stopnia skorelowania tych zmiennych, na przykładzie kilku zbiorów testowych. Badanie wartości *WKO* wraz z odpowiadającymi im wartościami współczynnika korelacji liniowej będzie **pierwszym sposobem** oceniania czy uwzględniane zmienne tworzą strukturę skupień. Wartości obu współczynników tj. dla korelacji odległościowej (4.1) i dla korelacji liniowej (4.2) są ze sobą powiązane. Jeżeli zmienne są skorelowane i nie tworzą struktury skupień, to powodują, że korelacja odległościowa jest dodatnia tym wyższa im współczynnik korelacji liniowej bliższy 1 lub -1. Jeżeli zmienne tworzą strukturę skupień, to w konsekwencji, powodują, że współczynnik korelacji liniowej pomiędzy każdą parą zmiennych ma często wartość istotnie różniąca się od zera. Na przykład jeśli dwie zmienne o



Rys. 4.1 Przykłady zbiorów obiektów dwuwymiarowych mających dodatnią wartość *WKO*(*u,v*). Po lewej obiekty tworzące strukturę skupień, po prawej zbiór obiektów bez struktury skupień. Źródło: Opracowanie własne.

wartościach rzeczywistych tworzą strukturę kilku skupień, to trudno tak ułożyć skupienia na płaszczyźnie by współczynnik korelacji liniowej pomiędzy obiema zmiennymi był bliski zeru. W szczególności, zgodnie z przyjętym sposobem generowania skupień (por. 1.4.1), założywszy, że każde dwa sąsiednie skupienia na obu zmiennych zachodzą na siebie w takim samym stopniu, mamy niewielkie szanse na to by małą liczbę skupień rozmieścić tak na płaszczyźnie, by otrzymać współczynnik korelacji liniowej bliski zeru. Takie rozmieszczenie byłoby łatwiejsze gdyby liczba skupień była większa od kilku. Istotne są w tym kontekście liczebności skupień oraz ich kształt.

Drugim ze sposobów rozstrzygania o tym czy zbiory zmiennych są istotne dla struktury skupień może być badanie wzajemnych relacji pomiędzy wartościami *WKO* w zależności od liczby zmiennych użytych do obliczania jego. W przypadku badania *WKO* pomiędzy dwiema zmiennymi możemy porównywać jej wartość tylko z wartością uogólnionego współczynnika korelacji liniowej. W przypadku większej od 2 liczby zmiennych użytych do obliczania *WKO* możemy badać również, na przykład :

- zmianę $WKO((u,v),s)$ w porównaniu z $WKO(u,v)$, $WKO(u,s)$, $WKO(s,v)$;
- zmianę $WKO((u,v,t),s)$ w porównaniu z $WKO((u,v),s)$, $WKO((u,t),s)$, $WKO((t,v),s)$;
- zmianę $WKO((u,v,t),s)$ w porównaniu z $WKO((u,v),(s,t))$.

Wymienione zostały tylko niektóre kombinacje możliwych porównań, te, które wydają się naturalne. Oczywiście, można dowolnie modyfikować podzbiory porównywanych zmiennych pamiętając o tym, żeby tak dobierać zbiory zmiennych by z porównań można było wyciągnąć logiczne wnioski. Zbadanie korelacji odległościowej pomiędzy wszystkimi możliwymi podzbiorami danego zbioru zmiennych jest bardzo czasochłonne, gdyż obliczanie *WKO* dla 20 par obserwacji musimy wielokrotnie powtarzać w celu znalezienia stabilnej wartości średniej. Chcąc wykorzystać korelację odległościową do wybierania zmiennych należy zatem pamiętać o tym, że już dla 10 zmiennych obliczenie wszystkich możliwych korelacji odległościowych mija się z sensem. Ogólnie rzecz biorąc idea badania wzajemnych relacji wartości *WKO* polega na tym by znajdywać zmienne, które dołączone do jakiegoś zbioru innych zmiennych powodują wzrost wartości *WKO* nie będąc skorelowanymi ze zmiennymi z tego zbioru. Wówczas przyczyną wzrostu wartości musi być istotność sumy mnogościowej obu zbiorów zmiennych dla struktury skupień. Niech ilustracją będzie przykład następujących zbiorów testowych.

Stosując pierwszy sposób używać będziemy symbol $WKL(u,v)$, który oznacza oszacowanie punktowe współczynnika korelacji liniowej pomiędzy zmiennymi u,v obliczone w podobny sposób jak *WKO*(*u,v*) tzn. w postaci średniej arytmetycznej współczynnika korelacji liniowej dla szeregu *l* obserwacji obu zmiennych

$$WKL(u,v) = \frac{\frac{1}{l}\sum_{i=1}^{l} u_i v_i - \bar{u}\,\bar{v}}{s_u{}^u s_v}. \qquad (4.2)$$

Tabele 4.1 i 4.2 przedstawiają wartości *WKO* oraz współczynnika korelacji liniowej będące średnimi arytmetycznymi z 200 wartości współczynnika obliczonym na 30 parach obiektów (losowano po 30 obiektów do każdego obiektu zbioru) ze zbiorów testowych Z1 i Z2. Zbiór testowy Z1 składa się ze 200 obiektów opisanych przez 6 zmiennych. Obserwacje na pierwszych trzech zmiennych zostały wygenerowane z mieszaniny 4 trójwymiarowych rozkładów normalnych o jednostkowej macierzy kowariancji, po 50 obserwacji z każdego

rozkładu. Obserwacje mierzone na tych trzech zmiennych tworzą dość wyraźną strukturę 4 skupień (*overlap*=0,1 por. 1.4.1). Obserwacje na trzech ostatnich zmiennych zostały wygenerowane z trójwymiarowego rozkładu normalnego o zerowym wektorze średnich i macierzy kowariancji mającej 1 na przekątnej oraz 0,5 poza przekątną. Trzy ostatnie zmienne są nieistotne dla struktury skupień istniejącej w zbiorze i są parami dość silnie skorelowane liniowo. Zbiór Z2 składa się z 200 obiektów, które na trzech pierwszych współrzędnych są identyczne z trzema pierwszymi współrzędnymi obiektów zbioru Z1, zaś na ostatnich trzech współrzędnych zostały wygenerowane z trójwymiarowego rozkładu normalnego o zerowym wektorze średnich i jednostkowej macierzy kowariancji. Trzy ostatnie zmienne są nieistotne dla struktury skupień istniejącej w zbiorze i są parami nieskorelowane.

Zauważmy na początku, ze wyciąganie wniosków o istotności zmiennych dla struktury skupień na podstawie *WKO* jak również współczynnika korelacji liniowej ma sens tylko wtedy gdy ich wartości są wyższe (w przypadku skorelowania liniowego co do wartości bezwzględnej) od pewnego dodatniego progu. Przyjęcie takiego założenia jest konieczne, gdyż pomiędzy korelacjami bliskimi zeru mogą zachodzić dowolne, przypadkowe zależności. Jak wynika z liczb zawartych w tablicy 4.1, jedną wspólną wartość **progu** dla *WKO* oraz wartości bezwzględnej *WKL* najbezpieczniej jest przyjąć na poziomie ***pr*1=0,15.** Poniżej tej wartości niekiedy możemy wnioskować jedynie o nieistotności zmiennych dla struktury skupień.

Tabela 4.1 Współczynniki korelacji liniowej $WKL(u,v)$ pomiędzy parami zmiennych ze zbioru Z1.

	$WKL(u,v)$				
	v=1	v=2	v=3	v=4	v=5
u=2	0,110				
u=3	0,754	0,294			
u=4	0,007	0,051	0,026		
u=5	0,145	0,001	0,093	0,458	
u=6	0,006	0,029	0,025	0,483	0,394

Źródło: obliczenia własne.

W tabeli 4.1 wszystkie pary zmiennych składające się z jednej zmiennej istotnej i jednej nieistotnej mają *WKO* bliski zeru. W tabeli 4.2 wszystkie pary

zmiennych nieistotnych lub pary składające się z jednej zmiennej istotnej i jednej nieistotnej mają *WKO* bliski zeru. Możemy zatem zanotować następującą własność.

Tabela 4.2 Współczynniki korelacji odległościowej $WKO(u,v)$ pomiędzy parami zmiennych ze zbioru Z1.

	$WKO(u,v)$				
	v=1	v=2	v=3	v=4	v=5
u=2	0,283				
u=3	0,540	0,281			
u=4	0,003	0,043	0,030		
u=5	0,046	0,004	0,002	0,259	
u=6	0,000	0,015	0,020	0,220	0,185

Źródło: obliczenia własne.

Własność 1. Jeżeli dwie zmienne są nieistotne dla struktury skupień oraz są nieskorelowane lub jedna zmienna jest istotna a druga nie, to *WKO* pomiędzy tymi zmiennymi jest bliski zeru.

Tabela 4.3 Współczynniki korelacji liniowej $WKL(u,v)$ pomiędzy parami zmiennych ze zbioru Z2.

	$WKL(u,v)$				
	v=1	v=2	v=3	v=4	v=5
u=2	0,110				
u=3	0,754	0,294			
u=4	0,007	0,051	0,026		
u=5	0,145	0,001	0,093	0,458	
u=6	0,006	0,029	0,025	0,483	0,394

Źródło: obliczenia własne.

Tabela 4.4 Współczynniki korelacji odległościowej $WKO(u,v)$ pomiędzy parami zmiennych ze zbioru Z2.

	$WKO(u,v)$				
	v=1	*v*=2	*v*=3	*v*=4	*v*=5
u=2	0,283				
u=3	0,540	0,281			
u=4	0,003	0,043	0,030		
u=5	0,046	0,004	0,002	0,259	
u=6	0,000	0,015	0,020	0,220	0,185

Źródło: obliczenia własne.

Biorąc pod uwagę wartość *WKO* pomiędzy skorelowanymi liniowo zmiennymi nieistotnymi zbioru Z1 (zmienne 3, 4 ,5) można zauważyć, że jest ona równa około 50-60% wartości współczynnika korelacji liniowej. W przypadku zmiennych istotnych dla struktury skupień relacja ta może być znacznie wyższa – wartość *WKO* jest podnoszona ponad poziom 60% przez strukturę skupień. Na przykład, wartość *WKO* dla pary zmiennych (2,3), prawie dorównująca wartości współczynnika korelacji liniowej jest znacznie wyższa od 60% wartości współczynnika korelacji liniowej. Podobnie, w przypadku pary (1,2) *WKO* przewyższa znacznie przewyższa znacznie wartość *WKL* choć ta nie przekracza progu *pr1*. Zanotujmy zatem następującą własność, w której zaproponowane zostały wartości graniczne odnoszące się do relacji pomiędzy korelacją odległościową a liniową:

Własność 2.

Jeżeli $WKO(u,v) > 0{,}8 * |WKL(u,v)| \ \wedge \ |WKL(u,v)| > 0{,}25$, to wynika stąd, że para zmiennych (u,v) jest istotna dla struktury skupień.

Własność 3.

Jeżeli $WKO(u,v) > |WKL(u,v)| \ \wedge \ WKO(u,v) > 0{,}2$, to wynika stąd, że para zmiennych (u,v) jest istotna dla struktury skupień.

Tabela 4.5 Wartości $WKO((u,v),s)$ pomiędzy parami zmiennych a pojedynczymi zmiennymi ze zbioru Z1.

u	*v*	*s*=1	*s*=2	*s*=3	*s*=4	*s*=5	*s*=6
1	2			0,479	0,032	0,042	0,026
1	3		0,320		0,013	0,021	-0,006
1	4		0,143	0,363		0,226	0,176
1	5		0,128	0,306	0,186		0,151
1	6		0,151	0,338	0,163	0,172	
2	3	0,484			0,047	0,001	0,011
2	4	0,131		0,154		0,192	0,183
2	5	0,158		0,107	0,206		0,163
2	6	0,150		0,133	0,184	0,140	
3	4	0,340	0,186			0,190	0,155
3	5	0,352	0,157		0,213		0,144
3	6	0,368	0,183		0,185	0,143	
4	5	0,015	0,006	0,011			0,255
4	6	0,015	0,027	0,024		0,282	
5	6	0,023	-0,006	-0,016	0,302		

Źródło: obliczenia własne.

Im wyższa wartość bezwzględna współczynnika korelacji liniowej tym większa pewność uzyskanych wniosków (do momentu gdy *WKL* nie jest zbyt bliska jedności patrz komentarz niżej), dlatego też próg dla *WKL* został podwyższony do 0,25. Z kolei, przy wyższych wymaganiach dla stosunku *WKO* do *WKL*, można tak jak we własności 3, obniżyć trochę próg dla współczynnika korelacji liniowej. Przy wartości *WKL* niewiele przekraczającej *pr*1 łatwiejsze jest, na przykład, uzyskanie współczynnika korelacji odległościowej większego od $0{,}8*|WKL(u,v)|$. Wynika to stąd, że przy słabo skorelowanych zmiennych obserwacje będące ich realizacjami dają wartości $WKL(u,v)$ o dużo większym względnym odchyleniu od wartości współczynnika korelacji liniowej pomiędzy zmiennymi u,v, których realizacjami są te obserwacje. Podobnie jak w

przypadku korelacji pomiędzy parami zmiennych, patrząc na odpowiednio wysokie relacje *WKO* pomiędzy wartością dla pary zmiennych a wartością dla *WKO* pomiędzy parą a jedną zmienną (por. tablica 4.3), możemy wnioskować, że fakt ten musi być spowodowany istotnością tych trzech zmiennych dla struktury skupień. Na przykład $WKO((1,3),2) = 0{,}320$ jest większe od średniej arytmetycznej $WKO(1,2) = 0{,}283$ i $WKO(3,2) = 0{,}281$ oraz znacznie przewyższa średnią arytmetyczną korelacji liniowej $WKL(1,2) = 0{,}110$ i $WKL(3,2) = 0{,}294$. W przypadku *WKO* dla trójek zmiennych nieistotnych również może zdarzyć się, że analizowane relacje będą podobne ale nigdy $WKO((u,v),s)$ nie przekracza średniej arytmetycznej z dwóch współczynników korelacji liniowej $WKL(u,s)$ i $WKL(v,s)$. To, po prostu, nie jest możliwe gdy dodatni *WKO* jest konsekwencją tylko i wyłącznie skorelowania zmiennych bez udziału istotności tych zmiennych dla struktury skupień. Podobnie jak w przypadku par zmiennych można by pokusić się o ustalenie progu powyżej którego przyrost *WKO* spowodowany dołączeniem trzeciej zmiennej do pary zmiennych świadczy o istotności otrzymanej trójki zmiennych dla struktury skupień.

Wszystkie powyższe rozważania, które dotyczą wyciągania wniosków z odniesienia do siebie wartości *WKO* oraz *WKL* mają sens tylko dla wartości *WKL* świadczących o istnieniu związku statystycznego pomiędzy dwiema zmiennymi. Wartości te nie mogą być bliskie +/-1 ponieważ świadczyłyby wówczas o związku funkcyjnym (lub prawie funkcyjnym) a w takim przypadku wartości *WKO* byłyby również odpowiednio wysokie. Te wysokie wartości *WKO* mogłyby być konsekwencją związku funkcyjnego pomiędzy dwiema zmiennymi i nie muszą świadczyć o istnieniu struktury skupień w zbiorze danych. Bezpiecznym progiem granicznym dla $|WKL|$, do którego ma sens porównywanie wartości *WKO* oraz *WKL* i stosowanie własności 2 i 3 może być liczba **0,8**. Zastrzeżenie tego typu jest związane z rozróżnianiem pojęcia zmiennej zbędnej (*redundant*) do pojęcia zmiennej nieistotnej dla struktury skupień. Jeśli związek pomiędzy dwiema zmiennymi jest funkcyjny, to jedną z nich należy uznać za zbędną. Na obecnym etapie konstruowania metody nie wykluczamy z rozważań zmiennych zbędnych.

Odnotujmy jeszcze jeden bardzo istotny fakt dotyczący zachowania się wartości *WKO* w przypadku gdy oba zbiory składają się z kilku zmiennych. $WKO((1,2),(3,4)) = 0{,}145 < 0{,}151 = WKO((1,2),(3,4,5))$. A zatem dołączenie zmiennej nieistotnej dla struktury skupień do jednego ze zbiorów pomiędzy którymi obliczamy *WKO* nie oznacza, że otrzymamy niższą wartość współczynnika. Ten fakt zwany jest brakiem monotoniczności danego

wskaźnika i jest zasadniczą przeszkodą utrudniającą stosowanie wskaźników do selekcji zmiennych w zbiorach danych z dużą liczbą zmiennych.

Pierwszy sposób wykrywania istotności zmiennych dla struktury skupień można zastosować bezpośrednio tylko w przypadku małej liczby zmiennych. W przypadku np. kilkunastu lub większej liczby zmiennych, na ogół, wtedy gdy wszystkie zmienne przyczyniają się do utworzenia struktury skupień, wkład każdej pary zmiennych lub nawet trójki, będzie tak mały, że analizując ten podzbiór nie da się wyśledzić żadnej struktury skupień. W szczególności, przy przyjętym sposobie generowania struktur skupień (por. § 1.4), różnice parametrów położenia skupień są tak małe, że wykrycie struktury skupień nie jest możliwe poprzez analizę par zmiennych.

Przy drugim sposobie wykrywania zmiennych istotnych dla struktury skupień należy sprawdzić jak licznych podzbiorów zmiennych należy używać w celu wykrycia struktury skupień. Ponadto, istotne w tym kontekście jest również w jakich konfiguracjach tj. wzajemnym stosunku liczebności obu zbiorów *A, B* zmiennych (por. wzór (4.1)) należy badać wartości *WKO* by jak najlepiej wykrywać struktury skupień. W tabeli 4.6 przedstawiona jest analiza wartości *WKO*(*A,B*) w zależności od czterech czynników: stopnia zachodzenia skupień na siebie, liczby skupień, korelacji wewnątrzklasowej i liczebności zbiorów *A, B*. Wartości *WKO* są średnimi arytmetycznymi z 10 zbiorów każdego typu. Można zaobserwować prawidłowość polegającą na tym, że wartości *WKO* pomiędzy dwoma zbiorami zmiennych rosną wraz ze wzrostem liczby skupień – wówczas korelacja liniowa lepiej wychwytuje istnienie struktury. Ta prawidłowość kończy się jednak, gdy liczba skupień jest tak duża, że trudno odróżnić strukturę skupień od rozkładu równomiernego, ponieważ dla 8 skupień wartości są często niższe niż dla 4 lub 6 skupień. Z liczb zawartych w tabeli 4.6 można jeszcze wyciągnąć ważne wnioski dotyczące zachowania się *WKO*(*A,B*) w zależności od liczebności zbiorów *A, B*. Gdy liczebności zbiorów rosną, to wartości *WKO*(*A,B*) maleją. Gdy pełny zbiór zmiennych, tworzących strukturę skupień składa się z 6 zmiennych, to *WKO* pomiędzy parami zmiennych ma bardzo małe wartości, mniejsze od 0,05. Niewiele zmienia nawet istnienie korelacji wewnątrz skupień – jedynie w przypadku małej liczby skupień wartości są nieco wyższe. Istotne w tym kontekście jest również to, czy oba zbiory *A, B* wyczerpują w swojej sumie mnogościowej zbiór wszystkich zmiennych. Gdy dwa zbiory nie dopełniają się do zbioru wszystkich zmiennych, to *WKO* pomiędzy nimi może zachowywać się bardziej przypadkowo. Dokładniejsze badanie zachowania się *WKO* raczej nie ma sensu ze względu na silne uwarunkowane konfiguracją skupień (ich kształtem, względną liczebnością, ewentualną korelacją wewnątrzklasową). Istotne jest też to, że

Tabela 4.6 Analiza wartości *WKO*(*A,B*) w zależności od stopnia zachodzenia skupień na siebie, liczby skupień, korelacji wewnątrzklasowej i liczebności zbiorów *A, B*.

Overlap	Liczba skupień	Średnie wartości *WKO*(*A,B*) spośród 10 zbiorów danego rodzaju, w których strukturę skupień tworzy 6 zmiennych			
		Zbiory bez korelacji wewnątrzklasowej		Zbiory z korelacją wewnątrzklasową	
		$\lvert A\rvert = 2, \lvert B\rvert = 1$	$\lvert A\rvert = 2, \lvert B\rvert = 2$	$\lvert A\rvert = 2, \lvert B\rvert = 1$	$\lvert A\rvert = 2, \lvert B\rvert = 2$
0	2	0,564	0,039	0,601	0,041
	3	0,551	0,039	0,531	0,038
	4	0,476	0,035	0,434	0,030
	6	0,365	0,026	0,351	0,025
	8	0,262	0,021	0,297	0,020
0,1	2	0,021	0,002	0,248	0,017
	3	0,123	0,009	0,260	0,015
	4	0,214	0,016	0,155	0,013
	6	0,247	0,018	0,218	0,016
	8	0,235	0,018	0,230	0,017
0,2	2	0,020	0,001	0,170	0,010
	3	0,126	0,007	0,270	0,017
	4	0,218	0,014	0,224	0,016
	6	0,210	0,016	0,268	0,016
	8	0,325	0,017	0,218	0,017
0,3	2	-0,002	0,000	0,170	0,010
	3	0,071	0,005	0,270	0,017
	4	0,133	0,009	0,224	0,016
	6	0,192	0,014	0,268	0,016
	8	0,178	0,012	0,218	0,017
0,4	2	0,005	0,000	0,182	0,013
	3	0,071	0,006	0,209	0,015
	4	0,137	0,009	0,151	0,011
	6	0,123	0,014	0,194	0,015
	8	0,207	0,015	0,187	0,014

Źródło: opracowanie własne

czasochłonność obliczeń przy większej od kilku liczbie zmiennych będzie czynnikiem znacznie ograniczającym stosowalność *WKO*(*A,B*) gdy liczebności obu zbiorów *A, B* są wyższe od kilku.

Niewątpliwie użycie *WKO* daje wiele korzyści w porównaniu z innymi metodami służącymi do wybierania zmiennych istotnych dla struktury skupień. Najważniejsze to uniezależnienie się od konieczności grupowania obiektów zbioru danych oraz od konieczności znajomości liczby skupień. Dotychczas proponowane metody wykrywania struktury skupień możemy próbować zastąpić porównaniem korelacji odległościowej z liniową (pierwszy sposób) oraz porównaniem zmian wartości korelacji odległościowej wraz ze zmieniającą się liczbą zmiennych branych pod uwagę przy jej obliczaniu (drugi sposób). Cechy obu technik zestawiono w tabeli 4.7.

Na podstawie cech korelacji odległościowej przedstawionych w tabeli 4.8 można wnioskować, że pewną wadą korzystania z korelacji odległościowej jest ograniczenie stosowalności w przypadku dużej liczby zmiennych. Zauważmy jednak, że nie jest to wada dotycząca wszystkich możliwych aspektów zastosowań korelacji odległościowej. Na przykład, jeśli podzielimy zbiór wszystkich zmiennych na rozłączne podzbiory zmiennych parami skorelowanych ze sobą wewnątrz każdego zbioru, to na tym etapie możemy w dalszym ciągu (lub po raz pierwszy) stosować korelację odległościową do

Tabela 4.7 Zestawienie cech dwóch sposobów wykorzystania korelacji odległościowej do wykrywania zmiennych tworzących strukturę skupień.

Pierwszy sposób: porównanie *WKO* z *WKL* dla par zmiennych	Drugi sposób: porównanie *WKO* dla różnych podzbiorów zmiennych
1. Konieczność dokładnego ocenienia siły korelacji pomiędzy zmiennymi co jest możliwe tylko na silnych skalach pomiarowych. 2. Możliwość bezpośredniego użycia tylko przy małej liczbie zmiennych. 3. Niepotrzebna znajomość liczby skupień. 4. Nie ma potrzeby grupowania danych. 5. Ograniczenia liczbowe (dolne i górne) dla wartości *WKL* , dla których ten sposób można stosować.	1. Ograniczenia natury obliczeniowej, przy dużej liczbie zmiennych, przez konieczność badania obszernych podzbiorów zbioru wszystkich zmiennych. 2. Niepotrzebna znajomość liczby skupień. 3. Nie ma potrzeby grupowania danych.

Źródło: opracowanie własne

do oceniania istotności każdego z podzbiorów poprzez badanie wartości *WKO* pomiędzy dwiema dopełniającymi się częściami każdego z podzbiorów. Przedstawione w dalszej części monografii metody pozwolą ocenić czy dany

Tabela 4.8 Podstawowe cechy korelacji odległościowej w kontekście użycia jej wykrywania zmiennych tworzących strukturę skupień.

1. *WKO* można obliczać dla podzbiorów zmiennych o różnych liczebnościach choć najatrakcyjniejsze jest analizowanie wartości *WKO* dla par zmiennych, bo łatwiej wtedy wnioskować o istotności pojedynczych zmiennych.
2. *WKO* dla par zmiennych ma najlepsze własności dyskryminacyjne gdy liczba skupień jest umiarkowana; dla 2 skupień i dla większej od 6 liczby skupień *WKO* często traci te zdolności.
3. *WKO* nie posiada własności monotoniczności względem liczby zmiennych tzn. dołączenie np. zmiennej nieistotnej do jednego ze zbiorów zmiennych pomiędzy którymi obliczono *WKO* nie musi wiązać się ze spadkiem wartości *WKO*.
4. Wysoka wartość *WKO* nie musi świadczyć o istnieniu struktury skupień, może być konsekwencją skorelowania zmiennych.

Źródło: opracowanie własne.

podzbiór zmiennych parami skorelowanych ze sobą tworzy strukturę skupień. Zaproponowane zostaną również metody pozwalające na podzielenie zbioru wszystkich zmiennych na rozłączne podzbiory zmiennych spójnych w sensie ich istotności dla ewentualnej struktury skupień. W tych metodach ciężar grupowania zmiennych w spójne podzbiory będzie spoczywał zarówno na *WKO* jak i różnych formach współczynników korelacji. Zauważmy, że współczynniki korelacji (np. współczynnik korelacji liniowej), przy niejednoznacznościach wniosków wynikających z ich stosowania, mają jedną zasadniczą zaletę: jeśli zmienne są skorelowane, to można to wyśledzić ograniczając się do badania par zmiennych.

Gdy zmiennych jest dużo, to proponowane progi powyżej których oceniamy wartość *WKO* jako wysoką nie są już aktualne. Co więcej, wtedy nie można podać nowych wartości progów, gdyż wartości *WKO* są bardzo biskie zeru niezależnie od tego czy podzbiory zmiennych tworzą strukturę skupień czy nie. Dla 25 zmiennych (a nawet większej liczby zmiennych) oraz *overlap*=0,1 można jeszcze zauważyć wartość $WKO(A,B) \approx 0{,}1$ w przypadku gdy zbiory *A, B* „wyczerpują" zbiór wszystkich 25 zmiennych tworzących strukturę skupień tzn. gdy $|A|+|B|=25$. Lecz gdy zbiory *A, B* są wybrane mniej precyzyjnie, to *WKO* ma wartość bardzo bliską zeru. Ta własność została ustalona na podstawie badania kilku struktur skupień o takich konfiguracjach liczby zmiennych

istotnych oraz stopnia zachodzenia skupień na siebie. Struktury skupień tworzone przez zmienne w liczbie równej 50 lub wyższej (por. tab. 1.3) nie pozwalają na to by je wykryć przy pomocy podstawowej formy *WKO* pomiędzy dowolnymi dwoma rozłącznymi podzbiorami zmiennych tworzących strukturę skupień.

4.2. Wymuszanie wyższej wartości korelacji odległościowej

Wartości *WKO* pomiędzy parami zmiennych tracą swoją zdolność dyskryminacyjną w sensie wykrywania zmiennych istotnych dla struktury skupień wraz ze wzrostem liczby wszystkich zmiennych. Ten fakt wynika w prosty sposób z wartości różnic pomiędzy centrami sąsiednich skupień przedstawionymi w tablicy 1.3. Jak zauważono w tamtym paragrafie struktury skupień na małej liczbie wymiarów istotnych dla struktury stają się nierozróżnialne. Nawet w przypadku zbioru z np. 6 zmiennymi może zdarzyć się, że *WKO* dla niektórych par zmiennych jest bliskie zeru i nic z takiej wartości nie można wywnioskować. Gorsza jest również zdolność dyskryminacyjna korelacji odległościowej w przypadku najmniejszej tj. równej 2, liczby skupień. Możemy jednak znacznie poprawić zdolność dyskryminacyjną *WKO*, na przykład, ograniczając się do obliczania jego wartości tylko dla par zmiennych z obszarów większego zagęszczenia zbioru. Kierujemy się następującym rozumowaniem: jeżeli wartość *WKO* będzie obliczana w oparciu o pary obserwacji z obszarów o większym zagęszczeniu, to w przypadku istnienia struktury skupień te obszary określają skupienia w związku z czym przynależność obserwacji do różnych skupień może dać efekt wyższej wartości *WKO*. W ten sposób próbujemy wymuszać wyższą wartość *WKO*. Takie podejście będzie bardzo pomocne jeśli będziemy mogli dzięki niemu odróżnić czy dodatnia wartość *WKO* obliczona dla wszystkich obserwacji była konsekwencją skorelowania zmiennych czy istnienia struktury skupień. Jeśli dodatnia wartość *WKO* była konsekwencją skorelowania zmiennych, to ponowne obliczenie *WKO* na zagęszczonym „centrum” zbioru nie powinno w znacznym stopniu zmienić wartości *WKO*.

Technik powierzchniowo-gęstościowych wyszukiwania obszarów zbioru o większym zagęszczeniu można znaleźć w literaturze wiele. Przykładem może być metoda zamieszczona w książce Gordona (1999). Metoda ta ma charakter sekwencyjny. Polega na kolejnym wybieraniu obserwacji, które mają najmniejszą sumę odległości do np. piątego najbliższego sąsiada. Obserwacje wybierane są sekwencyjnie spośród tych, które jeszcze nie zostały wybrane. A zatem, w miarę wybierania kolejnych obserwacji rozrzedzają się obszary gęste i zaczynają być wybierane obserwacje z obszarów, które na początku miały

mniejsze zagęszczenie. Parametrem determinującym rezultat tej metody jest numer najbliższego sąsiada. Tym parametrem można łatwo sterować w zależności od liczebności zbioru danych. Dla zbiorów składających się z kilkuset obiektów z pewnością dobra wartością jest numer 5. Dla zbiorów składających się z kilku tysięcy obiektów można tę liczbę zwiększyć do np. 20 (powinno to wystarczyć dla wykrycia zarówno kilku lub nawet kilkunastu skupień zarówno mniejszych jak i większych). Drugim parametrem jest odsetek liczby obiektów, które zostaną wybrane. Metodę tę oznaczmy symbolem $S(k,\beta)$ gdzie k oznacza liczbę najbliższych sąsiadów każdego obiektu (w wspomnianych rozważaniach Gordona 5), zaś β oznacza odsetek liczby obiektów, które zostały wybrane w tej metodzie o charakterze sekwencyjnym (np. połowa).

Zaproponujmy inną metodę wybierania obszarów zbioru o większym zagęszczeniu obserwacji. Rozważmy np. 5% najmniejszych odległości pomiędzy parami obserwacji. Zanotujmy dla każdej obserwacji i ze zbioru liczbę $lb(i)$ mówiącą o tym ile razy ta obserwacja występowała wśród wszystkich rozważanych par. Uporządkujmy obserwacje malejąco według liczby $lb(i)$. Chcąc uwzględnić, na przykład, najbardziej zagęszczoną połowę zbioru uwzględniamy tylko pierwszą połowę obserwacji w uporządkowanym ciągu $lb(i)$. Ta technika różni się tym od techniki Gordona, że wybór obszaru jest zdeterminowany przez część najmniejszych odległości (5%) nie zaś przez wskazanie liczby najbliższych sąsiadów. Metodę tę oznaczmy symbolem $G(\alpha,\beta)$ gdzie α oznacza odsetek par z najmniejszymi odległościami (w powyższych rozważaniach 5%), β oznacza odsetek liczby obiektów o najwyższych wartościach $lb(i)$ (w powyższych rozważaniach połowa).

Zupełnie innym w charakterze sposobem wymuszania większej (ogólniej – o zadanej wartości) korelacji odległościowej jest badanie jej wartości przez ograniczenie się tylko do tych wartości, które zostały obliczone dla wskazanych ciągów l par obserwacji (por. def. 4.1). Możemy narzucić warunek by, na przykład, ograniczyć się do wyznaczenia wartości *WKO* jako średniej arytmetycznej tylko tych spośród kilkuset (lub n) powtórzeń, dla których wartość była dodatnia. Takie postępowanie ma na celu zbadanie stabilności wartości *WKO* i jest wariantem badania zróżnicowania wartości *WKO* , z których obliczana jest średnia. Techniką tę również możemy zastosować w celu rozstrzygnięcia czy dodatnia wartość *WKO* obliczona dla wszystkich obserwacji była konsekwencją skorelowania zmiennych czy istnienia struktury skupień. Kierujemy się następującym rozumowaniem. Jeżeli wartość *WKO* wyraźnie większa od zera jest konsekwencją skorelowania zmiennych, to uśrednianie

Tabela 4.9 Wartości *WKO* przy różnych rodzajach wymuszeń gęstościowych wyższej wartości *WKO*. . Cztery pierwsze zmienne tworzą strukturę skupień, cztery ostatnie są parami silnie skorelowane *WKL*=0,75.

Pary zmiennych	*WKO* dla całego zbioru danych	*G*(5%,50%)	*G*(5%,75%)	*G*(10%,50%	*G*(10%,75%	*S*(5,50%)
1, 2	0,34	0,30	0,59	0,44	0,39	0,36
1, 3	0,12	0,11	0,08	0,17	0,05	0,12
1, 4	0,42	0,42	0,42	0,42	0,52	0,38
2, 3	0,39	0,42	0,43	0,37	0,49	0,38
2, 4	0,27	0,25	0,32	0,27	0,22	0,29
3, 4	0,21	0,28	0,19	0,28	0,26	0,22
5, 6	0,50	0,51	0,44	0,46	0,53	0,51
5, 7	0,42	0,41	0,40	0,44	0,44	0,40
5, 8	0,39	0,37	0,44	0,44	0,45	0,39
6, 7	0,56	0,60	0,49	0,51	0,50	0,53
6, 8	0,49	0,48	0,53	0,55	0,51	0,49
7, 8	0,52	0,55	0,49	0,48	0,59	0,50

Pominięte pary zmiennych miały wartości w całym wierszu mniejsze od 0,15.
Źródło: obliczenia własne.

tylko dodatnich wartości nie zmieni znacząco końcowej wartości, bo tylko nieliczne spośród kilkuset ciągów dawały ujemną wartość *WKO*. Jeżeli zaś wartość *WKO* wyraźnie większa od zera jest konsekwencją istnienia struktury skupień, to szereg kilkuset wartości *WKO* jest, na ogół, o wiele bardziej zróżnicowany i uśrednienie tylko wartości dodatnich może wyraźnie zwiększyć wartość końcową. Rozumując podobnie możemy narzucić warunek obliczania *WKO* przez uśrednienie tylko tych wartości, dla których współczynnik korelacji liniowej spełnił określony warunek (np. był dodatni/ujemny). Postępując w ten sposób kierujemy się następującą logiką. Jeżeli dodatnia wartość *WKO* była konsekwencją skorelowania zmiennych, to wartości *WKO* będą podążały dokładnie śladem wartości *WKL* i obliczanie *WKO* przez uśrednianie tylko tych wartości, którym odpowiadają dodatnie wartości *WKL* nie zmieni znacząco końcowej wartości *WKO*. Jeśli zaś dodatnia wartość *WKO* była konsekwencją istnienia struktury skupień, to dodatnie wartości *WKL* nie mają tak ścisłego związku z odpowiadającymi im wartościami *WKO*, w wobec czego te ostatnie mogą, po uśrednieniu, dać inną wartość końcową.

Tabela 4.10 Wartości *WKO* przy różnych rodzajach wymuszeń mechanicznych wyższej wartości *WKO*. Cztery pierwsze zmienne tworzą strukturę skupień, cztery ostatnie są parami silnie skorelowane *WKL*=0,75.

Pary zmiennych	*WKO* dla całego zbioru danych	*WKO* dla 50% obiektów najbliższych jednemu z krańcowych obiektów	*WKO* dla 50% obiektów najbliższych drugiemu z krańcowych obiektów	*WKO* dla 50% obiektów mających najmniejsze wartości pierwszej zmiennej pary	*WKO* dla 50% obiektów mających najmniejsze wartości pierwszej zmiennej pary
1, 2	./.	0,32	0,18	0,40	./.
1, 3	0,12	0,19	0,08	0,17	0,32
1, 4	./.	0,47	0,21	./.	0,22
2, 3	0,29	./.	0,45	0,37	0,18
2, 4	0,27	0,25	0,32	0,17	0,25
3, 4	0,21	./.	0,19	0,28	0,16
5, 6	0,53	0,51	0,44	0,46	0,29
5, 7	0,42	0,37	0,38	0,40	0,34
5, 8	0,37	0,37	0,29	0,44	0,28
6, 7	0,49	0,40	0,49	0,32	0,42
6, 8	0,45	0,41	0,33	0,35	0,38
7, 8	0,50	0,41	0,40	0,38	0,34

Symbol ./. oznacza wartość mniejszą od 0,15. Pominięte pary zmiennych miały wartości w całym wierszu mniejsze od 0,15.

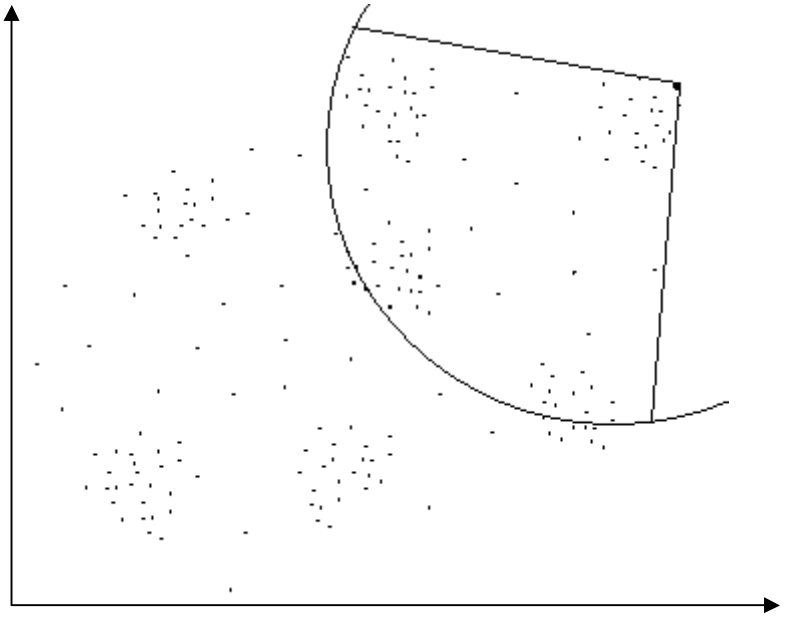
Źródło: obliczenia własne.

Dla przykładu rozważmy zbiór *n*=200 obiektów opisanych przez 8 zmiennych. Cztery pierwsze zmienne tworzą dość wyraźną (*overlap*=0,1) strukturę 4 skupień o jednakowych liczebnościach, zaś cztery ostatnie zmienne są silnie skorelowanymi (*WKL*=0,75) parami zmiennymi o standaryzowanym rozkładzie normalnym.

Liczy zawarte w tabeli 4.9 pozwalają na wyciągnięcie następujących wniosków:

- wymuszenia metodą $S(k,\beta)$ (ostatnia kolumna) dają bardzo stabilne wartości *WKO* niewiele wyższe od wartości podstawowej; wzrosty te mają jednak bardzo podobny charakter dla par zmiennych istotnych oraz dla nieistotnych ale skorelowanych;
- najbardziej wyraźne wzrosty wartości *WKO* daje technika $G(\alpha,\beta)$, ale niestety podobnie, wzrosty te mają jednak bardzo podobny charakter dla par

zmiennych istotnych oraz dla nieistotnych ale skorelowanych; dla par zmiennych istotnych wzrosty są znacznie bardziej niestabilne tzn. w niektórych wariantach są znaczne, a w niektórych nie są nawet wzrostami lecz spadkami.



Rys. 4.2 Zbiór punktów ze strukturą siedmiu skupień. Obszar zaznaczony wycinkiem koła, to połowa zbioru najbliższa jednemu z punktów krańcowych.

Źródło: Opracowanie własne.

Kolejną techniką wymuszania wyższej wartości korelacji odległościowej, bardzo pożyteczną w przypadku nieco większej liczby skupień może być technika wyboru podzbioru całego obszaru zmienności dwóch cech, na którym będzie obliczany *WKO*. Jak widać z liczb zawartych w tabeli 4.6 korelacja odległościowa traci swą zdolność dyskryminacyjną wraz ze wzrostem liczby skupień do stopnia, w którym trudno odróżnić strukturę skupień od rozkładu równomiernego. Można zaradzić tej wadzie wybierając obszar na którym będzie obliczany *WKO* ale nie w sensie obszaru o większym zagęszczeniu obserwacji. Obszar ten można wybrać mechanicznie, na przykład jedną połowę obserwacji określoną przez wartość średnią jednej ze zmiennych z rozważanej pary, lub obszar centralny zakresu zmienności cechy. W przypadku skali pomiarowej, na której nie ma dobrej średniej można posłużyć się łatwo kryterium połowy odległości od ustalonego (np. krańcowego obiektu) lub ¼ maksymalnej odległości od obiektu równooddalonego od dwóch krańcowych obiektów zbioru (w przypadku słabszych skal pomiarowych). Po wybraniu takiego obszaru liczba

skupień znacznie się zmniejszy czego konsekwencją powinny być znacznie wyższe wartości *WKO*. Ilustracją tej idei jest rysunek 4.2. Przez obiekty krańcowe rozumiemy dwa obiekty o największej odległości w zbiorze wszystkich odległości pomiędzy parami obiektów. Obliczanie *WKO* na zbiorze zaznaczonym przez wycinek koła może dać znaczny wzrost w przypadku w zbiorze istnieje struktura skupień, ponieważ na mniejszej liczbie skupień

Tabela 4.11 Techniki wymuszania wyższej/innej wartości *WKO* stosowane w dalszej części pracy.

Oznaczenie techniki	Definicja techniki
1. *G*(5%,50%)	Ograniczenie zbioru, na którym obliczany jest *WKO* do zbioru 50% wszystkich obiektów występujących na początku ciągu wszystkich obiektów uporządkowanych malejąco względem częstości występowania obiektów wśród 5% par z najmniejszymi odległościami.
2. *G*(5%,75%)	Ograniczenie zbioru, na którym obliczany jest *WKO* do zbioru 75% wszystkich obiektów występujących na początku ciągu wszystkich obiektów uporządkowanych malejąco względem częstości występowania obiektów wśród 5% par z najmniejszymi odległościami.
3. *G*(10%,50%)	Ograniczenie zbioru, na którym obliczany jest *WKO* do zbioru 50% wszystkich obiektów występujących na początku ciągu wszystkich obiektów uporządkowanych malejąco względem częstości występowania obiektów wśród 10% par z najmniejszymi odległościami.
4. *G*(10%,75%)	Ograniczenie zbioru, na którym obliczany jest *WKO* do zbioru 75% wszystkich obiektów występujących na początku ciągu wszystkich obiektów uporządkowanych malejąco względem częstości występowania obiektów wśród 10% par z najmniejszymi odległościami.
5. *S*(5,50%)	Ograniczenie zbioru, na którym obliczany jest *WKO* do zbioru 50% wszystkich obiektów wybieranych sekwencyjnie. W każdym kroku wybierany jest obiekt, który ma najmniejszą sumę odległości do 5 najbliższych sąsiadów i obiekt ten usuwany jest ze zbioru przed wyborem obiektu w kroku następnym.
6. Pkr1	Ograniczenie zbioru, na którym obliczany jest *WKO* do zbioru połowy wszystkich obiektów, najbliższych jednemu z dwóch obiektów krańcowych zbioru.
7. Pkr2	Ograniczenie zbioru, na którym obliczany jest *WKO* do zbioru połowy wszystkich obiektów, najbliższych drugiemu z dwóch obiektów krańcowych zbioru.

Źródło: opracowanie własne.

korelacja odległościowa ma lepsze zdolności dyskryminacyjne. W przypadku gdy para zmiennych nie posiada struktury skupień wybór mechaniczny jakiegokolwiek spójnego, dużego obszaru z całego zakresu zmienności nie powinien wpłynąć na wartość *WKO* obliczaną na tym obszarze.

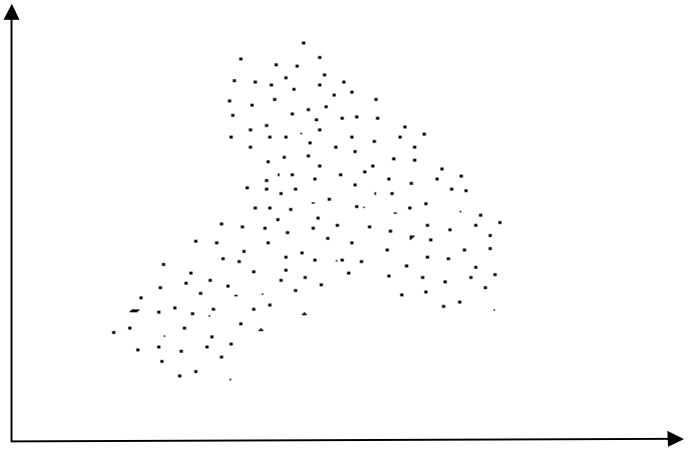
Tabela 4.10 zawiera przykładowe zmiany *WKO* otrzymane dla zbioru n=200 obiektów opisanych przez 8 zmiennych, dla czterech rodzajów wymuszeń mechanicznych. Cztery pierwsze zmienne tworzą dość wyraźną (*overlap*=0,1) strukturę 8 skupień o jednakowych liczebnościach, zaś cztery ostatnie zmienne są silnie skorelowanymi (*WKL*=0,75) parami zmiennymi o standaryzowanym rozkładzie normalnym. Z liczb zawartych w tej tablicy wynika, że metoda mechanicznego wybierania podzbioru zbioru danych może dać bardzo dobre efekty dla wzrostu wartości *WKO* w przypadku liczby skupień większej od kilku, struktury trudnej do rozróżnienia od rozkładu równomiernego.

W niektórych przypadkach, z wartości nieznaczącej (poniżej 0,15) *WKO* wzrasta do wartości rzędu 0,4. Mankamentem jest jednak to, że wartość *WKO* dla par zmiennych nieistotnych ale skorelowanych, również wzrasta choć w znacznie mniejszym stopniu. Ponadto wadą jest to, że wzrosty wartości *WKO* w przypadku par zmiennych istotnych występują dość rzadko. Przeprowadzono badanie na zestawie PROB zbiorów (por. 1.4.1) z wyraźnymi i prostymi strukturami skupień. Zbadano podział dwuwymiarowych (dla każdej pary zmiennych) obszarów zmienności na dwie połowy, określone przez zbiory obiektów najbliższych każdemu z punktów krańcowych (tak jak w 3 i 4 kolumnie tabeli 4.10). Ponad 30% wzrosty wartości *WKO* (dla całego zbioru mającego wartość wyższą od 0,3) zaobserwowano tylko w przypadku 17% zbiorów. Ten wynik jednak wystarcza do tego, by poprawić efektywność korelacji odległościowej, gdyż wzrosty miały miejsce głównie w przypadku dużej liczby (6) skupień.

Należy zaznaczyć, że ocenienie wielkości zmian wartości *WKO* przy pomocy przedstawionych technik jest łatwiejsze gdy wartość *WKO* dla całego zbioru jest umiarkowanie wysoka, tj. wyższa od 0,3 ale niezbyt bliska 1. Wówczas możemy posługiwać się kryterium względnego wzrostu. W przypadku małych wartości *WKO* ocenianie zmian i porównywanie zmian dla różnych par zmiennych jest bardziej kłopotliwe.

Wszystkie zbadane techniki wymuszania wyższej/innej wartości *WKO* są zdefiniowane w tabeli 4.11. Stosując korelację odległościową oraz techniki wymuszania innej/wyższej jej wartości można skonstruować bardzo efektywną dwustopniową metodę selekcji zmiennych. Zadaniem realizowanym w pierwszym stopniu jest pogrupowanie zmiennych na dwa rozłączne podzbiory na podstawie wartości *WKO* oraz *WKL* dla par zmiennych (bądź liczniejszych podzbiorów zmiennych). W drugim stopniu wystarczy rozstrzygnąć który z podzbiorów ma silniejsze cechy świadczące o istnieniu struktury skupień przy

pomocy kryterium większej zmiany wartości *WKL*. Taka metoda osiągnęła bardzo wysokie wskaźniki efektywności na zestawie PODST zbiorów danych generowanych – pamięć i precyzja były powyżej 95%. Problemem pozostaje jednak to czy duża zmiana wartości *WKL* na różnych podzbiorach zbioru danych



Rys. 4.3 Przykład dwuwymiarowego zbioru obiektów bez struktury skupień, z diametralnie różnymi wartościami *WKL* na obu częściach.

jest cechą wystarczającą do tego by stwierdzić istnienie struktury skupień w całym zbiorze. Przykład przedstawiony na rysunku 4.3 pokazuje, że odpowiedź na to pytanie jest raczej negatywna. Z drugiej jednak strony, w problemach empirycznych nie spotyka się zmiennych o tak wyszukanych postaciach rozkładów. Wymuszanie wyższej wartości *WKO* zostanie wykorzystane w pierwszym etapie algorytmu selekcji do grupowania zmiennych w homogeniczne podzbiory zmiennych. Przez $WKO^7(u,v)$ oznaczać będziemy najwyższą z siedmiu wartości *WKO* odpowiadających siedmiu technikom wymuszeń z tabeli 4.11.

4.3. Kilkustopniowe grupowanie obiektów w dwa skupienia

Jak zaznaczono w poprzednim paragrafie, przy założeniu, że zbiory zmiennych tworzących różne struktury skupień są rozłączne, podział zbioru wszystkich zmiennych na rozłączne podzbiory zmiennych skorelowanych w sensie *WKO* oraz *WKL* można uzyskać w oparciu o wartości tych współczynników dla par zmiennych (lub liczniejszych podzbiorów gdy

zmiennych jest dużo). Pozostaje zatem rozstrzygnąć który z podzbiorów ma silniejsze cechy świadczące o istnieniu struktury skupień.

Można zaproponować metodę grupowania obiektów (na każdym z wyróżnionych podzbiorów zmiennych) taką, która pozwoli wykrywać ewentualne struktury skupień bez znajomości liczby skupień – zakładając tylko, że jest ona niewielka. Idea takiej metody polega na kilkustopniowym sprawdzaniu czy zbiór można pogrupować w dwa skupienia. Kilkustopniowość należy rozumieć w tym sensie, że grupujemy cały zbiór danych na dwa skupienia i oceniamy ich separowalność (miary zostaną zaproponowane w dalszym ciągu paragrafu). Następnie, każde z dwóch skupień ponownie dzielimy na dwa skupienia i oceniamy ich separowalność itd. Okazuje się, że jeżeli liczba skupień jest niewielka, to wystarczą trzy stopnie takiej weryfikacji, na ogół nawet dwa. Po takim kilkustopniowym grupowaniu miara ogólnej separowalności skupień powinna być uzależniona od separowalności wyodrębnionych skupień na tym samym stopniu jak i na różnych stopniach w celu uniknięcia sytuacji, w której dwa skupienia uzyskane w drugim stopniu są bardzo dobrze separowalne ale żadne z nich nie jest wyraźnie oddzielone od drugiego skupienia uzyskanego w pierwszym stopniu, więc i tak są ze sobą „połączone".

Miarę separowalności dwóch skupień można łatwo skonstruować wykorzystując w tym celu miary stabilności podziału oraz miary podobieństwa dwóch podziałów tego samego zbioru (por. § 1.2). Na przykład, jeżeli zbiór danych został podzielony na dwa skupienia, to podział części składającej się z połowy obiektów z pierwszego skupienia najbliższych drugiemu oraz połowy obiektów z drugiego skupienia najbliższych pierwszemu powinien być bardzo podobny do podziału uzyskanego przy dzieleniu całego zbioru. W przeciwnym przypadku świadczyłoby to o tym, że przy dzieleniu połowy zbioru najbliższej obszarowi rozgraniczenia dwóch skupień otrzymujemy inny wynik podziału, który kwestionuje rozgraniczenie dwóch skupień otrzymanych przy dzieleniu całego zbioru. Miarą podobieństwa dwóch podziałów będzie zmodyfikowany indeks Randa. W celu ocenienia jakości podziału pierwszego, całego zbioru, można też wykorzystać stabilność tego podziału zarówno w odniesieniu do całego zbioru jak i, na przykład połowy zbioru najbliższej obszarowi rozgraniczenia. Miarą stabilności może być analiza replikacji lub miara oparta na indeksie sylwetkowym. W ten sposób dość łatwo jest zdyskredytować podział zbioru obiektów opisanych przez zmienne silnie skorelowane, gdyż stabilność podziału obszaru rozgraniczenia (połowy zbioru) będzie dużo słabsza od stabilności podziału całego zbioru. Pozostaje ustalić najlepszą z tych miar i wybrać najlepszy jej wariant. W celu uniknięcia konieczności wprowadzania parametrów określających którą część zbioru tj. jak dużą i jak położoną, wybrać w celu zweryfikowania jakości podziału zbioru, zauważmy, że na ogół

otrzymujemy podział na dwa skupienia o różnych liczebnościach. Wobec tego, najbardziej naturalnym wyjściem z sytuacji wydaje się wzięcie jako obszaru rozgraniczenia całego mniejszego skupienia i części większego. W ten sposób rozwiążemy problem wyboru obszaru rozgraniczenia. Pozostaje ustalić jaką część większego skupienia włączyć do obszaru rozgraniczenia. Należy wziąć pod uwagę zarówno aspekt czasu obliczeń (istotny w eksperymentach symulacyjnych) jak i ograniczenia liczby obiektów z większego skupienia do stopnia, który różniłby się znacznie od całego zbioru. Biorąc pod uwagę to, że większe skupienia jest, na ogół, znacznie większe od mniejszego oraz to, że jakość podziału na dwa skupienia najłatwiej oceniać gdy są one mniej więcej równoliczne, zaproponujmy 1/3 część większego skupienia.

Wobec tego jakość podziału zbioru lub, w następnych stopniach jednego z wyróżnionych skupień, na dwa skupienia będziemy weryfikować przez porównanie tego podziału z podziałem podzbioru, który nazywać będziemy obszarem granicznym, składającego się ze wszystkich obiektów wchodzących w skład mniejszego skupienia i 1/3 obiektów najbliższych mniejszemu skupieniu, wchodzących w skład większego skupienia. Przy pomocy kilku przykładowych zbiorów danych ustalimy czy miarą jakości będzie stabilność podziału obszaru granicznego na dwa skupienia czy też zgodność podziału obszaru granicznego na dwa skupienia z podziałem całego zbioru na dwa skupienia. Zaznaczmy, że na samym wstępie rezygnujemy z oceniania jakości w sposób polecany w literaturze tzn. przez badanie stabilności podziału całego zbioru na dwa skupienia. Takie podejście nie ma sensu, gdyż zbiory opisane przez zmienne

Tabela 4.12 Wartości współczynnika Randa podobieństwa podziału obszaru granicznego z podziałem całego zbioru na dwa skupienia oraz stabilność podziału obszaru granicznego na dwa skupienia.

Liczba skupień	2	3	4	6	8
Podobieństwo podziału obszaru granicznego z podziałem całego zbioru	0,05	0,81	0,90	0,79	0,93
Stabilność podziału na 2 skupienia	0,59	1,00	0,97	0,42	0,09

Źródło: obliczenia własne.

silnie skorelowane zawsze mają wysoką stabilność podziału na dwa skupienia a nie tworzą żadnej struktury skupień (por. tabl. 4.13). Rozważmy przykładowo 10 zbiorów danych generowanych, w których dwie zmienne tworzą dość wyraźną (*overlap*=0,1 oraz *overlap*=0,2) strukturę równolicznych 2, 3, 4, 6 oraz 8 skupień bez korelacji wewnątrz skupień (por. 1.4.1). W tabeli 4.12

przedstawione są średnie (z 2 zbiorów odpowiadających dwóm różnym wartościom *overlap*) wartości współczynnika Randa podobieństwa (na obszarze granicznym) podziału obszaru granicznego na dwa skupienia z podziałem całego zbioru na dwa skupienia i stabilności (replikacja przy 20 powtórzeniach, por. § 1.2) podziału obszaru granicznego na dwa skupienia. Podziały przeprowadzone były przy pomocy metody *k*-średnich, dla *k*=2.

Z liczb zawartych w tabeli 4.12 wynika, że zarówno indeks podobieństwa podziału jak i stabilność podziału mają podobne wartości, przy czym podobieństwo podziału ma najmniejsze dla dwóch skupień natomiast stabilność podziału na dwa skupienia ma mniejszą wartość dla 2 skupień a jeszcze mniejszą dla większych liczb skupień. Z przykładów zamieszczonych w tabeli 4.12 można by wyciągnąć wniosek o tym, że, przy odpowiednich ograniczeniach, obie te miary nadają się do oceniania jakości podziału zbioru na dwa skupienia. Tak jednak nie jest, ponieważ przykładowe wartości obu miar dla kilku wybranych konfiguracji rozkładów jednomodalnych (por. tab. 4.13) zdecydowanie pokazują, że indeks Randa podobieństwa podziałów obszaru granicznego z podziałem całego zbioru jest znacznie lepszym narzędziem. Wartości tej miary zawarte w tabeli 4.13 są zazwyczaj bardzo bliskie zeru, podczas gdy stabilność podziału na dwa skupienia jest wysoka nie tylko dla zmiennych jednomodalnych skorelowanych – również układy kilku zmiennych

Tabela 4.13 Wartości miary $R(2_1 \cup 1/3)$ oraz stabilności podziału na dwa skupienia dla wybranych zbiorów danych jednomodalnych.

Rodzaj rozkładu		Beta(1,1)	N(0,1) brak korelacji	N(0,1) słaba korelacja	N(0,1) średnio silna korelacja	N(0,1) silna korelacja
Dwie zmienne	$R(2_1 \cup 1/3)$	0,35	0,02	-0,03	0,10	0,01
	Stabilność podziału na 2 skupienia	0,33	0,34	0,56	0,54	0,66
Cztery zmienne	$R(2_1 \cup 1/3)$	-0,01	-0,01	0,06	0,14	-0,02
	Stabilność podziału na 2 skupienia	0,19	0,25	0,34	0,66	0,90

Źródło: obliczenia własne.

jednomodalnych nieskorelowanych wykazują stabilność podziału na dwa skupienia rzędu 0,30. Kierując się tymi spostrzeżeniami wybierzmy tę miarę jako miarę jakości podziału na dwa skupienia i oznaczmy ją symbolem $R(2_1 \cup 1/3)$. Należy zaznaczyć, że wartości miary $R(2_1 \cup 1/3)$ są silnie uzależnione od rodzaju struktury skupień, w szczególności od liczebności skupień. Jeżeli miara $R(2_1 \cup 1/3)$ będzie miała niską wartość nie znaczy to, że nie ma struktury skupień. Należy próbować podziału na dwa skupienia każdego ze skupień otrzymanych w pierwszym stopniu i zaproponować miarę siły struktury skupień uwzględniającą wskazania wynikające z podziałów wszystkich stopni. Taka miara zostanie zaproponowana w dalszym ciągu paragrafu. Wartości miary $R(2_1 \cup 1/3)$ w tabeli 4.13 są małe, na ogół bardzo małe, ale nie można poprzestać na porównaniu tylko tych wartości z analogicznymi z tabeli 4.12, gdyż dla niektórych struktur skupień liczby w tabeli 4.12 również są bliskie zeru.

W celu uzupełnienia ocen jakości podziału pierwszego stopnia na dwa skupienia należy każde z dwóch skupień podzielić na dwa skupienia. Te podziały będziemy nazywali podziałami drugiego stopnia. Dla każdego z dwóch podziałów drugiego stopnia również znajdujemy wartości miary $R(2_1 \cup 1/3)$. Podziały drugiego stopnia muszą zostać dokonane niezależnie od wskazań dla podziału pierwszego stopnia dlatego, że, jak widać w tabeli 4.12, w przypadkach niektórych struktur skupień, wartości miary $R(2_1 \cup 1/3)$ są zeru. Jeżeli istnieje struktura skupień, to wartości miary dla podziałów obu stopni są silnie uzależnione od liczebności, kształtu i rozkładów skupień ale, niezależnie od tych czynników, ostateczna miara siły struktury skupień, w najprostszej postaci może być największą ze wszystkich trzech wartości miary czyli

$$R(A) = \max\left\{R_1(2_1 \cup 1/3);\, R^*_{2,1}(2_1 \cup 1/3);\, R^*_{2,2}(2_1 \cup 1/3)\right\} \quad (4.3)$$

gdzie trzy miary w nawiasie odnoszą się do, odpowiednio, podziału pierwszego stopnia, podziału drugiego stopnia pierwszego skupienia (otrzymanego w pierwszym stopniu) i podziału drugiego stopnia drugiego skupienia, A – zbiór zmiennych opisujących obiekty. Dokładniej,

$$\begin{aligned} R^*_{2,1}(2_1 \cup 1/3) &= \min\left\{R_{2,1}(2_1 \cup 1/3); R_{2,1\cup(1,2)}(2_1 \cup 1/3)\right\} \\ R^*_{2,2}(2_1 \cup 1/3) &= \min\left\{R_{2,2}(2_1 \cup 1/3); R_{2,1\cup(1,1)}(2_1 \cup 1/3)\right\} \end{aligned} \quad (4.4)$$

gdzie indeks 2,1 odnosi się do wartości miary $R(2_1 \cup 1/3)$ dla podziału drugiego stopnia pierwszego skupienia (z pierwszego stopnia), symbol

$2,1 \cup (1,2)$ odnosi się do wartości miary dla zbioru obiektów w chodzących w skład mniejszego skupienia z dwóch na które podzielono (w drugim stopniu) pierwsze skupienie (z pierwszego stopnia) oraz wszystkich obiektów z drugiego skupienia (z pierwszego stopnia). Analogicznie, symbol $2,1 \cup (1,1)$ odnosi się do wartości miary dla zbioru obiektów w chodzących w skład mniejszego skupienia z dwóch na które podzielono (w drugim stopniu) drugie skupienie (z pierwszego stopnia) oraz wszystkich obiektów z pierwszego skupienia (z pierwszego stopnia), zaś indeks 2,2 odnosi się do wartości miary $R(2_1 \cup 1/3)$ dla podziału drugiego stopnia drugiego skupienia (z pierwszego stopnia). Wartość miary (4.3) będzie pomocna w wykrywaniu struktur skupień w liczbie większej od kilku, bo w tych przypadkach, miara $R_1(2_1 \cup 1/3)$ może mieć małą wartość. Do miary jakości podziału drugiego stopnia np. $R_{2,1}(2_1 \cup 1/3)$, dla pierwszego skupienia, została dołączona (w formie wyboru mniejszej z dwóch we wzorach (4.4)) miara separowalności mniejszego skupienia (z dwóch otrzymanych w podziałach drugiego stopnia) i całego innego skupienia z pierwszego stopnia, po to by nie zwiększać wartości miary $R_1(2_1 \cup 1/3)$ w sytuacji gdy część ustalonego skupienia (z pierwszego stopnia) jest „bliższa" drugiemu skupieniu niż drugiej części (z drugiego stopnia) ustalonego skupienia.

Z tabeli 4.12 wynika, że zaproponowana miara (4.3) spisuje się najsłabiej gdy są tylko dwa skupienia. Przyczyną tego jest to, że algorytm dzielenia zbioru, w przypadku większej od dwóch liczby skupień, ma większy wybór i zawsze dokonuje podziału w odniesieniu do najbardziej wyraźnego obszaru rozgraniczenia. Gdy jest tylko jeden obszar rozgraniczenia, czyli dwa skupienia to, algorytm musi się nim zadowolić (o ile w ogóle znajdzie ten obszar).

Zaproponujmy technikę pseudobootstrapową, która może trochę poprawić efektywność wyszukiwania obszaru rozgraniczającego dwa skupienia i, bardzo rzadko, obniża jakość oceny braku takiego obszaru w przypadku zmiennych o rozkładach jednomodalnych lub równomiernych (tzn. rzadko istotnie podnosi wartość miary (4.3)). Zadaniem techniki pseudobootstrapowej będzie zniwelowanie wspomnianej niedogodności, poprzez zniwelowanie przypadkowości rozkładu wartości zmiennych w obszarze rozgraniczenia. Zniwelowanie tej przypadkowości można uzyskać zastępując oryginalny zbiór obserwacji zbiorem bardzo podobnym, który powstaje poprzez zastąpienie każdego oryginalnego obiektu obiektem leżącym „blisko" oryginalnego. Zaproponujmy następującą formułę generowania jednej obserwacji pseudobootstrapowej $x_v^{'}$ dla obserwacji oryginalnej x_v (po przeprowadzonej uprzednio unitaryzacji zerowanej). Dla ustalonej zmiennej v :

$$x_v^{'} = x_v + 0,2 \cdot r \qquad (4.5)$$

gdzie r – liczba wylosowana z odcinka [0;1] (niezależnie dla każdej zmiennej v i każdego obiektu). Po unitaryzacji zerowanej wszystkie wartości każdej zmiennej v leżą na odcinku [0;1] więc formułę (4.5) można uznać za generowanie losowe obserwacji leżącej „blisko" obserwacji oryginalnej.

Tabela 4.14 Wartości współczynnika Randa podobieństwa podziału obszaru granicznego z podziałem całego zbioru na dwa skupienia i stabilności podziału obszaru granicznego na dwa skupienia (średnie z 20 zbiorów pseudobootstrapowych).

Liczba skupień	2	3	4	6	8
$R(2_1 \cup 1/3)$	0,18	0,59	0,91	0,66	0,43
Stabilność podziału na 2 skupienia	0,51	0,77	0,92	0,41	0,51

Źródło: obliczenia własne.

Tabela 4.15 Wartości miary $R(2_1 \cup 1/3)$ oraz stabilność podziału na dwa skupienia dla wybranych zbiorów danych jednomodalnych (średnie z 20 zbiorów pseudobootstrapowych).

Rodzaj rozkładu		Beta (1,1)	N(0,1) brak korelacji	N(0,1) słaba korelacja	N(0,1) średnio silna korelacja	N(0,1) silna korelacja
Dwie zmienne	$R(2_1 \cup 1/3)$	-0,03	-0,01	0,01	0,11	-0,02
	Stabilność podziału na 2 skupienia	0,23	0,11	0,61	0,83	0,72
Cztery zmienne	$R(2_1 \cup 1/3)$	-0,00	0,01	0,17	-0,00	0,01
	Stabilność podziału na 2 skupienia	0,22	0,74	0,31	0,81	0,99

Źródło: obliczenia własne.

W celu sprawdzenia efektu generowania nowych obserwacji, w tabelach 4.14 i 4.15, przedstawiono wartości miar stabilności i miary $R(2_1 \cup 1/3)$ po zastąpieniu każdej obserwacji oryginalnej nową obserwacją zgodnie z formułą (4.5) dla tych samych zbiorów, których charakterystyki są zawarte w tabelach 4.12 i 4.13. Porównując tabele 4.12 i 4.14, można zauważyć, że zasadniczo rzecz biorąc, wartości niewiele różnią się od siebie, ale w przypadku dwóch skupień dołączone obserwacje pseudobootstrapowe poprawiają nieco wartość indeksu

Randa. Porównując tabele 4.13 i 4.15 można zauważyć, że wartości indeksu Randa poprawiły się znacznie. W porównaniu z tabelą 4.13, w której, w dwóch przypadkach wartości są wyraźnie wyższe od zera, dla zbiorów pseudobootstrapowych stają się bliskie zeru. Można zatem liczyć na to, że użycie zbiorów pseudobootstrapowych pozwoli efektywniej dyskredytować kombinacje zmiennych (zwłaszcza skorelowanych) jednomodalnych pod kątem tworzenia przez nie struktury skupień. Wobec tego ostateczną miarą jakości podziału na dwa skupienia obiektów opisanych zmiennymi ze zbioru A będzie

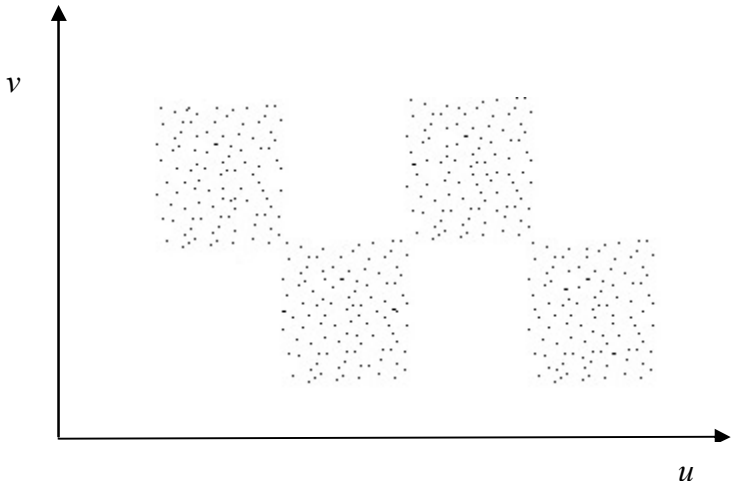
$$\max\left\{R(A), R\left(A^{boots}\right)\right\} \qquad (4.6)$$

gdzie $R\left(A^{boots}\right)$ oznacza średnią wartość miary (4.3) z 20 zbiorów pseudobootstrapowych dla zbioru zmiennych A.

4.4. Sformułowanie metody dla różnych rodzajów danych

4.4.1. Metoda dla zbiorów danych ze zmiennymi ciągłymi

Zmienne, które nie tworzą struktury skupień powinny mieć rozkład jednomodalny lub równomierny. Ta cecha, z formalnego punktu widzenia, nie jest jednak wystarczającym kryterium do wyszukiwania zmiennych istotnych dla struktury skupień, ponieważ może zdarzyć się, że zmienne istotne mają również



Rys. 4.4 Przykład zbioru punktów dwuwymiarowych ze strukturą skupień, która po zrzutowaniu na każdą z osi da rozkład równomierny.

Źródło: opracowanie własne.

rozkład jednomodalny lub nawet równomierny. W pracy Walesiaka (2005) przedstawiony jest przykład dwóch zmiennych dyskretnych, tworzących strukturę kilku skupień, takich, że każda z tych zmiennych ma rozkład równomierny. Inny przykład tego typu jest przedstawiony na rysunku 4.4. Oba te przykłady mają cechy wspólne, każde następne skupienie na każdej zmiennej zaczyna się tam gdzie ostatnie się kończy oraz wszystkie skupienia maja na każdej zmiennej rozkłady równomierne. Można oczywiście wyszukać przykład skupień o rozkładach nierównomiernych, „uzupełniających się" w ten sposób, by w efekcie, każdy rozkład brzegowy był równomierny, ale w praktyce zbiorów empirycznych występowanie takich rozkładów jest mało prawdopodobne. Na ogół, dąży się do wyeliminowania zmiennych o rozkładach równomiernych, gdyż nie mają one zdolności dyskryminujących (por. *Walesiak*, 2005). Taka eliminacja niesie jednak ze sobą pewne niebezpieczeństwo, zwłaszcza w przypadku większej od kilku liczby skupień, które po zrzutowaniu na jeden wymiar mogą dać rozkład o charakterze równomiernym.

W ogólnym przypadku, na początku algorytmu, rezygnujemy wobec tego z użycia wskaźników skupialności lub wielomodalności, stosowanych do pojedynczych zmiennych. Wskaźniki te zostaną ewentualnie wykorzystane w szczególnych przypadkach, w których skorelowanie zmiennych tworzących strukturę skupień (albo nie) jest bardzo wysokie i badanie *WKO* lub *WKL* nie ma dostatecznej zdolności dyskryminacyjnej. Wykorzystanie wskaźników skupialności w takich przypadkach, w odniesieniu do pojedynczych zmiennych ma logiczne uzasadnienie. Tym silniejsze jest to uzasadnienie im wyższe skorelowanie zmiennych tworzących strukturę skupień. Na początku algorytmu skoncentrujemy się na wykorzystaniu zdolności korelacji odległościowej do wyszukiwania zmiennych tworzących strukturę co najmniej dwóch skupień. Zaproponowany zostanie algorytm oparty na badaniu par zmiennych, który, siłą rzeczy, nie będzie potrafił wybrać jednoelementowego zbioru zmiennych tworzącego strukturę skupień, lecz w zamian za to będzie bardzo efektywny w selekcji większej od 1 liczby zmiennych tworzących strukturę skupień. Zauważmy, że najlepsze metody dotychczas skonstruowane również nie posiadają umiejętności wyszukiwania jednoelementowych zbiorów zmiennych tworzących struktury skupień (por. HINoV i jej modyfikacje).

Własności korelacji odległościowej przedstawione w poprzednich paragrafach sugerują skonstruowanie dwuetapowej metody wybierania zmiennych istotnych dla struktury skupień:

- w pierwszym etapie należy pogrupować wszystkie zmienne w rozłączne podzbiory wykorzystując do tego *WKO* oraz *WKL* dla par zmiennych.;
- w drugim etapie należy rozstrzygnąć czy każdy z wyodrębnionych podzbiorów zmiennych tworzy jakąś strukturę skupień w przypadku selekcji

bezwzględnej lub wybrać podzbiór zmiennych, który ma najsilniejsze wskaźniki istnienia struktury skupień w przypadku selekcji porównawczej.

W ten sposób otrzymamy metodę, która pozwoli niekiedy wykrywać kilka niezależnych od siebie struktur skupień, a ponadto, podjąć decyzję bezwzględną o tym ile jest struktur skupień w zbiorze danych i które zmienne którą strukturę tworzą. Jeżeli nie uda się rozstrzygnąć o istotności wyodrębnionych podzbiorów zmiennych w sposób bezwzględny, to, w uproszczonej wersji, otrzymamy decyzję porównawczą, która wybierze zbiór zmiennych tworzących strukturę skupień najsilniejszą w sensie zaproponowanych mierników.

Pierwszy etap metody polegający na pogrupowaniu zmiennych w podzbiory spójne pod względem ich istotności dla tworzonych przez nie struktur skupień obejmuje pierwsze 7 kroków algorytmu. Drugi etap realizowany jest w 8 i 9 kroku. Kolejne kroki w są następujące.

<u>Metoda bezwzględna *KO_ciągłe* selekcji zmiennych ciągłych</u>

Krok 1. Ustal i=1, $B_0 = VV$ (zbiór wszystkich zmiennych).

Krok 2. Znajdź dla wszystkich par zmiennych (u,v) ze zbioru B_{i-1}

$$R(u,v) = \frac{1}{2}\left(WKO^7(u,v) + \left|WKL^7(u,v)\right|\right).$$

Krok 3. Znajdź wśród liczb $R(u,v)$ znalezionych w kroku 2 najwyższą wartość. Parę odpowiadającą tej wartości oznaczmy przez (u_0, v_0). Jeśli $WKO(u_0, v_0) < 0{,}2$ oraz $WKL(u_0, v_0) < 0{,}2$, to idź do kroku 9.

Krok 4. Pogrupuj przy pomocy metody 2-średnich wszystkie zmienne ze zbioru B_{i-1} na dwa podzbiory: jeden przypisany do pary (u_0, v_0), oznaczmy ten zbiór przez A_i , drugi zbiór zmiennych pozostałych do analizy, oznaczmy go przez B_i . Zmienną dyskryminującą jest $\frac{1}{2}(R(s,v_0) + R(s,u_0))$, gdzie $s \in B_{i-1}$, $s \neq u_0, v_0$. Punkty startowe do metody 2-średnich to $R(u_0, v_0)$ oraz najmniejsza wartość $R(u,v)$ spośród wartości znalezionych w kroku 2.

Krok 5. Jeśli $card(B_i) \geq 3$, to połóż i=i+1 i idź do kroku 2.

Krok 6. Jeśli $card(B_i) = 2$, to połóż $A_{i+1} = B_i$ i idź do kroku 8.

Krok 7**.** Jeśli $card(B_i) = 1$, tj. jeśli pozostała tylko jedna zmienna s, to sprawdź czy należy ją dołączyć do zbioru A_i. Dołącz pozostałą zmienną do tego zbioru gdy wartość $\max\{WKO^7(s,v), |WKL(s,v)|\} > 0{,}3$ dla jakiejkolwiek zmiennej $v \in A_i$. Idź do kroku 8.

Krok 8. Dla wszystkich zbiorów A_i rozstrzygnij przy pomocy kryteriów bezwzględnych czy można zbiór A_i zmiennych uznać za zbiór tworzący strukturę skupień. Zakończ postępowanie.

Krok 9. Uznaj, że nie ma zmiennych, które tworzą strukturę skupień. Zakończ postępowanie.

W powyższym sformułowaniu symbol $WKO^7(u,v)$ oznacza najwyższą spośród 7 wartości *WKO* obliczanych metodami z tabeli 4.11 oraz $WKO(u,v)$ dla całego zbioru. Analogiczne rozumieć należy $WKL^7(u,v)$. Kryteria bezwzględne z kroku ósmego są określone następującymi warunkami:

1. Duże wartości *WKO* w stosunku do wartości *WKL* dla jakiejkolwiek pary zmiennych ze zbioru A_i, czyli:

$$WKO > 0{,}8 \cdot |WKL| \quad \wedge \quad 0{,}25 < |WKL| < 0{,}8 \text{ lub}$$

$$WKO > |WKL| \quad \wedge \quad 0{,}2 < |WKL| < 0{,}8 .$$

Jeżeli par spełniających jeden z tych warunków jest więcej niż jedna, to uwzględniamy wartości pary z najwyższym *WKO*.

2. Względnie duże wartości *WKO* przy bardzo małej wartości *WKL* dla jakiejkolwiek pary zmiennych ze zbioru A_i, czyli:

$$WKO > 0{,}2 \ \wedge \ |WKL| < 0{,}2 \text{ lub}$$

$$WKO > 0{,}25 \ \wedge \ |WKL| < 0{,}1 .$$

Jeżeli par spełniających jeden z tych warunków jest więcej niż jedna, to uwzględniamy wartości pary z najwyższym *WKO*.

3. Wartość miary (4.6) większa od 0,4.

4. Pozytywna ocena wizualna pary zmiennych (u_0, v_0), która była początkiem zbioru A_i jeżeli $WKL(u_0, v_0) > 0{,}8$.

<u>Metoda porównawcza *KO_ciągłe* selekcji zmiennych ciągłych</u>

Jeśli zadanie selekcji zmiennych ogranicza się do wybrania tylko jednego zbioru zmiennych tworzących strukturę skupień, którego istnienie zakładamy, i jeśli nie podejmiemy decyzji bezwzględnej odnośnie żadnego ze zbiorów A_i, to w kroku 8, spośród wszystkich zbiorów A_i należy wybrać podzbiór, który ma najsilniejsze charakterystyki liczbowe świadczące o tym, że zmienne z tego zbioru tworzą strukturę skupień czyli podzbiór, który ma najwyższą wartość miary (4.6).

Wartość miary (4.3) powyżej 0,4 to bardzo wysoka wartość, która powinna świadczyć o istnieniu struktury skupień, jak wynika z liczb zawartych w tabelach 4.13 i 4.15. Tylko w przypadku rozkładów skośnych miara (4.3) ma niekiedy wartość wyższą od 0,3. Oczywiście, wysokie wartości miary (4.3) mogą być również konsekwencją istnienia obserwacji nietypowych lub bardzo silnego skorelowania jednej pary zmiennych ze A_i. Zakładamy, że te dwa przypadki nie powinny mieć miejsca. Istnienie niewielkiej liczby obserwacji nietypowych, leżących relatywnie blisko siebie, które mogłyby dać dużą wartość miary (4.6), eliminujemy zerując wartość miary $R(A_i)$ wtedy gdy liczebność mniejszego skupienia występującego przy jej obliczaniu (por. (4.3)) jest mniejsza od 5% liczebności całego zbioru. To zerowanie ma miejsce zarówno przy obliczaniu wartości $R(A_i)$ oraz przy obliczaniu tej wartości dla 20 zbiorów pseudobootstrapowych (dla każdego z nich niezależnie). Bardzo wysokie (WKL>0,8) skorelowanie par zmiennych, które mogłyby dać dużą wartość miary (4.6), eliminujemy rezygnując w tym przypadku z oceny istnienia struktury skupień przy pomocy miary (4.6) i proponujemy ocenę wizualną, którą w przypadku pary zmiennych łatwo przeprowadzić. W eksperymentach symulacyjnych na dużej liczbie zbiorów trudno zastosować ocenę wizualną. W szczególności, w eksperymencie na zestawie PODST nie będzie ona stosowana. Takie podejście nie będzie miało złych konsekwencji, gdyż najsilniej skorelowane zmienne o rozkładach jednomodalnych (będące parami zmiennych, dla których WKL=0,75) nie powinny mieć empirycznego WKL>0,8 obliczanego za pomocą wzoru (4.2). Wobec tego żadne podzbiory zmiennych nieistotnych nie będą wymagały oceny wizualnej. To może być pewnym ułatwieniem, bo wśród zbiorów zmiennych tworzących wyraźną strukturę skupień wielokrotnie będą występować pary zmiennych z WKL>0,8 dla których wartość miary (4.6) będzie bardzo wysoka. Z dużą dozą prawdopodobieństwa można jednak założyć, że tak wyraźne struktury skupień zostałyby łatwo wykryte przy zastosowaniu jakiejkolwiek wizualizacji wartości dwóch zmiennych.

Wyjaśnienia wymaga również stosowanie zaproponowanych algorytmów w przypadku pojedynczych zbiorów empirycznych (por. rozdz. 5). Formalnie rzecz biorąc, zgodnie ze sformułowaniem algorytmu, wizualnie oceniać można tylko parę początkową każdego wyodrębnionego zbioru A_i zmiennych – na ogół, parę z najwyższą wartością WKL. Pierwszeństwo przed tą oceną powinna mieć miara (4.6) gdyż dotyczy ona całego zbioru A_i. Jak postąpić jednak gdy para początkowa nadaje się tylko do oceny wizualnej (ze względu na wysoki WKL), ale w zbiorze A_i są również pary zmiennych tworzące zbiór zmiennych nadający się do oceny przy pomocy miary (4.6) ? Ponadto, jeżeli zmiennych w

zbiorze A_i jest więcej niż kilka (czyli więcej niż w eksperymencie symulacyjnym), to czy nadal można polegać na spełnieniu kryteriów (1 oraz 2) decyzji bezwzględnej przez jedną parę zmiennych z kilkudziesięciu par? Wydaje się, że nie ma precyzyjnej odpowiedzi na te pytania. Na pewno odsetek poprawnych decyzji opartych na mierze (4.6) rośnie wraz ze wzrostem odsetka par zmiennych niezbyt silnie skorelowanych. Podobnie, stopień poprawnych decyzji opartych na kryteriach bezwzględnych 1 i 2, rośnie wraz ze wzrostem odsetka par spełniających te kryteria. Każdy zbiór empiryczny należy zbadać bardzo wnikliwie pod względem każdego z zaproponowanych kryteriów.

Dwa pierwsze kryteria decyzji bezwzględnej oparte na porównaniu *WKO* z *WKL*, w eksperymencie symulacyjnym, odnoszą się do wartości tych współczynników dla całego zbioru danych. Można te kryteria rozumieć w odniesieniu do połowy obiektów otrzymanych przy użyciu dowolnej metody z tabeli 4.11. Ta możliwość zostanie wykorzystana przy stosowaniu algorytmu *KO_ciągłe* do zbiorów empirycznych. W przypadku analizowania pojedynczych zbiorów jest możliwość dokładniejszej oceny ewentualnych wskazań przypadkowych i błędnych proponowanych mierników – tej możliwości nie ma w eksperymencie.

Powyżej zaproponowany algorytm porównawczy jest najprostszą wersją wykorzystującą ideę korelacji odległościowej oraz ideę wielokrotnego grupowania obiektów w dwa skupienia. Można przypuszczać, że w eksperymencie symulacyjnym na zestawie zbiorów PODST niektóre struktury o większej liczbie skupień zostaną ocenione nie dość dokładnie, ponieważ uwzględnione jest tylko grupowanie dwustopniowe. Jednak celem niniejszej monografii nie jest poszukiwanie algorytmu, który osiągałby optymalny wynik w eksperymencie symulacyjnym, ale raczej algorytmu ukierunkowanego na poprawną selekcję zmiennych w zastosowaniu do empirycznych zbiorów danych.

4.4.2. Metoda dla zbiorów danych ze zmiennymi porządkowymi

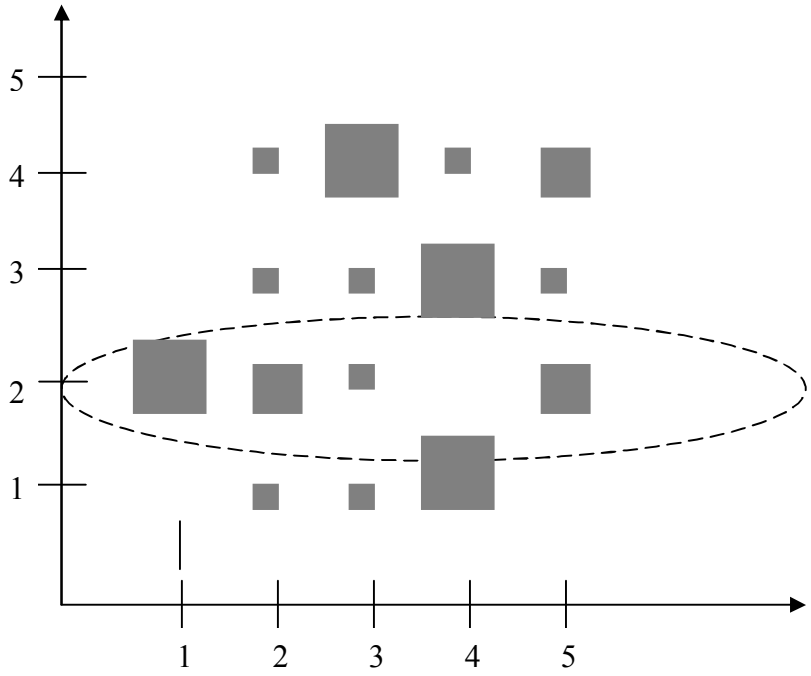
Dla zbiorów danych ze zmiennymi porządkowymi, do mierzenia siły związku dwóch zmiennych, zastosowano współczynnik korelacji liniowej dany wzorem (1.19). Jak wiadomo, wzór ten jest oparty na współczynniku korelacji tau-Kendalla. Na tym samym współczynniku oparta jest stosowana w tej monografii miara odległości GDM2 dla zmiennych porządkowych (por. *Gatnar, Walesiak,* 2004). Konsekwencją tych dwóch faktów jest to, że *WKL* dany wzorem (1.19) oraz *WKO* dany wzorem (4.1) przyjmują wartości bardzo silnie dodatnio skorelowane. Używanie obu współczynników w celu lepszego przeanalizowania związków pomiędzy skorelowaniem pojedynczych zmiennych

a istnieniem ewentualnej struktury skupień w rozkładzie tych zmiennych nie jest konieczne. Jako miarę, która będzie stosowana w analizie wybierzmy *WKL* dany wzorem (1.19). Zasadniczym powodem, który decyduje o takim wyborze było jest to, że *WKL* dokładniej niż *WKO* reaguje na pomijanie niektórych wartości zmiennych, co będzie analogią wymuszania innej wartości współczynników korelacji omawianego w przypadku zmiennych ciągłych (por. rys. 4.5).

Naturalnym podejściem do problemu wybierania zmiennych porządkowych tworzących strukturę co najmniej dwóch skupień jest ich łączenie w zbiory zmiennych skorelowanych ze sobą. Podobny punkt wyjścia do tego problemu zastosował Talavera (2000) konstruując metodę selekcji zmiennych nominalnych. Logika uzasadniająca takie podejście polega na spostrzeżeniu, że jeśli dwie zmienne o wartościach dyskretnych są skorelowane, to muszą tworzyć naturalną strukturę skupień wynikającą z wszystkich możliwych kombinacji wariantów obu zmiennych. Uwagę należy zwrócić na trudności wynikające z liczby wariantów. Jeżeli zmienna porządkowa ma dużo (więcej niż kilka) wariantów, to naturalna struktura skupień jaką ona tworzy będzie liczyła dużo skupień, a w parze z inną zmienną o podobnej liczbie wariantów, tych skupień może być kilkadziesiąt. Skorelowanie zmiennych kilkustanowych nie musi bowiem oznaczać tego, że przy rosnącej liczbie zmiennych liczba skupień jest cały czas taka sama i równa liczbie stanów (bądź największej z tych liczb) – liczba skupień może rosnąć, bo skorelowanie to jeszcze nie zależność funkcyjna. Wraz ze wzrostem liczby zmiennych może szybko (nawet wykładniczo) rosnąć liczba skupień w strukturze wynikającej z kombinacji wszystkich występujących wariantów dyskretnych. Taka struktura skupień nie jest dla nas interesująca ze względu na to, że mija się z ideą analizy skupień polegającą na ustaleniu ewentualnego istnienia niewielkiej liczby skupień.

Jeśli chcemy wykorzystać skorelowanie zmiennych porządkowych do ustalenia tego czy tworzą one strukturę kilku skupień, to musimy założyć, że liczba wariantów/stanów wszystkich zmiennych porządkowych opisujących obiekty jest rzędu kilku. Zauważmy, że przy tym założeniu, wartości *WKL* są czułe na wymuszenia o charakterze mechanicznym zakresu zmienności zmiennych, wtedy gdy obie zmienne tworzą strukturę kilku skupień. Wymuszenia zakresu zmienności o wiele łatwiej wprowadzić dla zmiennych porządkowych niż dla zmiennych ciągłych. Na przykład, w przypadku niewielkiej liczby stanów odpowiednikiem wybierania pewnego obszaru z dwuwymiarowej przestrzeni euklidesowej wartości zmiennych ciągłych jest narzucenie warunku o wyeliminowaniu jednego lub dwóch stanów ze wszystkich jakie przyjmuje dana zmienna porządkowa. Wyeliminowanie pewnych stanów jednej ze zmiennych, w oczywisty sposób, nie powinno mieć wpływu na wartość *WKL* gdy obie zmienne nie są skorelowane ze sobą.

Jeżeli dwie zmienne porządkowe tworzą strukturę kilku skupień i towarzyszy temu wysokie skorelowanie tych zmiennych, to eliminowanie jednego wariantu pociąga za sobą zmianę wartości *WKL,* ale zmiany te nie powinny być duże. Na przykład, jeśli założymy, że wykształcenie mierzone na skali pięciostanowej jest w pewnej populacji dość silnie skorelowane z zamożnością mierzoną również na skali pięciostanowej, to wyeliminowanie osób z wykształceniem najwyższym może obniżyć nieco wartość *WKL* ale wartość ta nadal pozostanie zdecydowanie dodatnia. Podobnie, wyeliminowanie osób najuboższych może spowodować (najprawdopodobniej) spadek wartości *WKL* ale nie powinien to być duży spadek. Ponadto, ewentualne spadki lub wzrosty wartości *WKL* spowodowane wyeliminowaniem jednego ze stanów powinny charakteryzować się niewielkimi różnicami w zależności od wyboru stanu.



Rys. 4.5 Dwie pięciowariantowe zmienne porządkowe tworzące strukturę kilku skupień (wielkość kwadratu jest proporcjonalna do liczebności skupienia).

Źródło: opracowanie własne.

Nieco większe mogą być zmiany wartości *WKL* wtedy gdy obie zmienne tworzą strukturę kilku skupień, ale nie towarzyszy temu wysoka wartość *WKL*. Jeżeli dwie zmienne tworzą strukturę skupień tego rodzaju, to wyeliminowanie

jednego z wariantów jednej zmiennej może wiązać się z wyeliminowaniem jednego skupienia lub większości obiektów z tego skupienia co może powodować, po pierwsze, dość istotną zmianę wartości *WKL*, po drugie, zmianę istotnie zróżnicowaną w zależności od wyboru wariantu. Tego rodzaju zależności są analogiczne do tych, które można było zauważyć przy wymuszaniu wyższej/innej wartości *WKO* dla zmiennych ciągłych. Zależności te zilustrowano na rysunku 4.5. Wartość *WKL* dla przedstawionego zbioru danych jest nieco większa od zera. Po usunięciu drugiego wariantu drugiej zmiennej (oś pionowa), wartość *WKL* może zmienić się zupełnie, nawet na ujemną. Logika powyżej przedstawionego rozumowania traci sens gdy liczba wariantów jest bardzo mała, na przykład wyeliminowanie jednego z dwóch wariantów uniemożliwia mierzenie siły korelacji.

Kierując się powyższymi rozważaniami skonstruowano metodę wybierania zmiennych istotnych dla struktury skupień, która oparta jest na wartości *WKL* oraz na zmianie wartości *WKL* w zależności od eliminowania poszczególnych wariantów pojedynczych zmiennych. Poniżej przedstawione są dwie wersje metody selekcji zmiennych porządkowych: bezwzględna i porównawcza. Liczbę wszystkich wariantów zmiennej v oznaczmy symbolem m_v.

Metoda bezwzględna *KO_porządkowe* selekcji zmiennych porządkowych

Krok 0. Połóż i=0 oraz zbiór wszystkich zmiennych na miejsce zbioru zmiennych pozostałych do analizy.

Krok 1. Połóż i=i+1. Dla wszystkich par zmiennych (ze zbioru zmiennych pozostałych do analizy) znajdź wartości $WKL(u,v)$. Ponadto, znajdź dla wszystkich par zmiennych (u,v) wartość $R(u,v)$ różnicy pomiędzy najwyższą a najniższą wartością *WKL* spośród $m_u + m_v$ wartości współczynnika *WKL* odpowiadających wyeliminowaniu poszczególnych wariantów obu zmiennych (za każdym razem eliminowany jest jeden wariant jednej zmiennej).

Krok 2. Znajdź wśród liczb $R(u,v)$ najwyższą wartość. Znajdź wśród liczb $WKL(u,v)$ najwyższą wartość. Parę odpowiadającą wyższej ze znalezionych dwóch wartości oznaczmy przez (u_0, v_0), gdy wyższa z dwóch wartości jest większa od 0,3. Jeśli nie, to zakończ postępowanie podejmując decyzję o braku zmiennych tworzących strukturę skupień.

Krok 3. Przypisz każdą zmienną s ze zbioru zmiennych pozostałych do analizy do zbioru A_i, przypisanego do pary (u_0, v_0), jeśli $\frac{1}{2}\left(\left|WKL(s,v_0)\right| + \left|WKL(u_0,s)\right|\right) > 0{,}3$. Przez B oznacz zbiór pozostałych zmiennych.

Krok 4. Jeśli pozostały co najmniej trzy zmienne w zbiorze B, to połóż na miejsce zbioru zmiennych pozostałych do analizy zbiór B i idź do kroku 1.

Krok 5. Jeśli pozostały dwie zmienne (u,v) w zbiorze B, to uznaj je za zmienne tworzące strukturę skupień jeżeli $R(u,v)>0{,}3$. Jeśli warunek ten nie jest spełniony, to uznaj tę parę za nieistotną. Idź do kroku 7.

Krok 6. Jeśli pozostała tylko jedna zmienna, to sprawdź czy można ją dołączyć do podzbioru zmiennych przypisanych do ostatniej pary (u_0, v_0). Dołącz pozostałą zmienną do ostatniego podzbioru zmiennych gdy wartość $|WKL|$ z którąkolwiek zmienną z pary (u_0, v_0) jest wyższa od 0,3.

Krok 7. Rozstrzygnij czy poszczególne, wyodrębnione zbiory A_i tworzą strukturę skupień za pomocą następującego kryterium: zbiór zmiennych tworzy strukturę skupień gdy średnia arytmetyczna $R(u,v)$ (czyli zmiany $WKL(u,v)$) ze wszystkich par (u,v) zmiennych z tego zbioru jest wyższa od 0,3.

<u>Metoda porównawcza *KO_porządkowe* selekcji zmiennych porządkowych</u>

Krok 0 do *Krok* 6 są takie same jak w metodzie bezwzględnej.

Krok 7. Spośród wszystkich wyodrębnionych zbiorów A_i, za zbiór zmiennych istotnych uznaj ten, który ma najwyższą średnią arytmetyczną $R(u,v)$ ze wszystkich par (u,v) zmiennych z tego zbioru.

Powyższe metody mogą być modyfikowane gdy pewne zmienne mają większą od kilku, ale mniejszą od 10, liczbę wariantów. Wówczas w kroku pierwszym można, na przykład, rozważać wszystkie wykluczenia pojedynczych wariantów oraz dwóch kolejnych stanów. W przypadku gdy liczba wszystkich wariantów jest większa od 10, to należy zwiększyć liczbę kolejnych, wykluczanych wariantów danej zmiennej. Uwzględnienie jako jednej z podstawowych miar zmiany *WKL* (a nie samej wartości *WKL*) powinno łagodzić złe skutki osłabienia zdolności skorelowania liniowego do rozstrzygania o istnieniu struktury skupień związane z tym, że przy rosnącej liczbie wzajemnie skorelowanych zmiennych może rosnąć liczba skupień. Na przykład, jeżeli na rysunku 4.5 struktura skupień byłaby rozłożona równomiernie pomiędzy wszystkie kombinacje wariantów obu zmiennych (byłaby to struktura 25 skupień), to zmienność *WKL* po usunięciu jednego wariantu jednej ze zmiennych byłaby bardzo niewielka. Zmienność *WKL* nie pomogłaby w ustaleniu istnienia takiej struktury skupień, ale takie struktury nie są dla nas interesujące, jak to zostało zaznaczone przed sformułowaniem algorytmu.

Obie zaproponowane metody zostaną sprawdzone na zestawie zbiorów zmiennych porządkowych (por. 1.4.2), ale należy zaznaczyć, że metody te mają niewielkie znaczenie praktyczne z uwagi na to, że rzadko spotyka się empiryczne zbiory danych, w których występowałyby tylko zmienne

porządkowe. Celem opracowania powyżej zaproponowanych metod jest pokazanie uniwersalnych możliwości korelacji odległościowej.

4.4.3. Metoda dla zbiorów danych ze zmiennymi binarnymi

Przeprowadzone zostały eksperymenty wstępne mające na celu zbadanie zachowania się *WKO* w przypadku gdy obiekty zbioru danych są opisane przez zmienne binarne. Zastosowano najpopularniejszą miarę odległości dla tego typu zmiennych tj. miarę Sokala-Michenera (por. tab. 1.1). W przypadku gdy badamy wartości *WKO* pomiędzy pojedynczymi zmiennymi, to $m=1$ (m – liczba zmiennych), wobec czego odległość jest równa 0 lub 1 w zależności od tego czy obiekty mają (odpowiednio) taką samą wartość na obu zmiennych czy różną. W pierwszym eksperymencie zbadano zachowanie się współczynnika korelacji odległościowej pomiędzy dwiema zmiennymi binarnymi bez struktury skupień.

Tabela 4.16 Wektory określające wartości zmiennych w strukturze sześciu skupień.

	$v1$	$v2$	$v3$	$v4$
Sk nr 1	1	0	0	1
Sk nr 2	1	1	1	1
Sk nr 3	1	0	1	0
Sk nr 4	0	1	0	1
Sk nr 5	0	0	0	1
Sk nr 6	0	1	1	0

Źródło: Brusco, 2004.

Zbadano wartość $WKO(u,v,l)$ przy $l=50$ tj. współczynnik korelacji odległościowej (4.1) dla 50-cio elementowych zbiorów par wylosowanych ze zbioru 1000 losowo wygenerowanych liczb ze zbioru $\{0, 1\}$. Losowanie 50-cio elementowych zbiorów par powtarzano 200 razy i $WKO(u,v)$ przyjęto równą średniej arytmetycznej z 200 powtórzeń. Ze 100 powtórzeń obliczania otrzymano maksymalną wartość $WKO(u,v)$ równą **0,021**.

W drugim eksperymencie zbadano zachowanie się współczynnika korelacji odległościowej pomiędzy zmiennymi binarnymi, z których jedna nie ma żadnej struktury skupień zaś druga zmienna jest jedną z czterech zmiennych opisujących strukturę 6 skupień daną wektorami przedstawionymi w tabeli 4.16. Wartości zmiennej bez struktury skupień losowane były z losowo wygenerowanego ciągu 1000 zer lub jedynek, zaś wartości zmiennej reprezentującej strukturę losowane były z ciągu 1200 czwórek, z których 200 pierwszych czwórek miało wartości takie jak w pierwszym skupieniu, 200 następnych czwórek miało wartości takie jak w drugim skupieniu, 200 kolejnych

czwórek miało wartości takie jak w trzecim skupieniu itd. Najpierw 50 razy losowano czwórkę następnie z czwórki wybierano wartość tej zmiennej (decydował numer zmiennej licząc od początku czwórki), która występowała w parze dla której *WKO* był obliczany. Drugą zmienną pary była zmienna bez struktury skupień, więc do otrzymanej liczby dołączano liczbę wylosowaną z ciągu 0 oraz 1. Podobnie jak w poprzednim eksperymencie, losowanie takich 50 par powtarzane było 200 razy. Podobnie jak w poprzednim eksperymencie, wartości *WKO* były bliskie zeru dla wszystkich czterech zmiennych z tabeli 4.16. Na 100 powtórzonych obliczeń maksymalna wartość $WKO(u,v)$ jaką otrzymano, to **0,027**.

Tabela 4.17 Zależność pomiędzy *WKO* a stopniem rozmycia skupień.

Stopień rozmycia skupień	0%	2%	4%	6%	8%	10%
Minimalna wartość *WKO*	0,10	0,091	0,088	0,080	0,076	0,062

Źródło: obliczenia własne.

W trzecim eksperymencie zbadano zachowanie się współczynnika korelacji odległościowej pomiędzy parami zmiennych z czterech zmiennych opisujących strukturę 6 skupień daną wektorami przedstawionymi w tabeli 4.16. Celem tego badania było przede wszystkim uzyskanie zależności pomiędzy stopniem rozmycia skupień a wartością *WKO*, sama wartość *WKO* dla poszczególnych par zmiennych nie była celem badania. Wyniki badania zawarte są w tabeli 4.17.

Minimalna wartość *WKO* jest rozumiana jako najmniejsza wartość *WKO* dla tych spośród sześciu par zmiennych, dla których wartość *WKO* przy zerowym rozmyciu skupień była wyższa od 0,1. Poszczególne wartości *WKO* dla wszystkich sześciu par były bardzo zróżnicowane, od wartości nawet ujemnych do 0,55. Celem tego eksperymentu jest zatem ustalenie jak wpływa zwiększanie stopnia rozmycia skupień na *WKO* ale tylko dla tych par, które przy zerowym rozmyciu skupień miały wartość *WKO* , którą można (w świetle poprzednich eksperymentów) uznać za wartość świadczącą o tym, że przyczyniają się one do tworzenia struktury skupień. Liczby przedstawione w tablicy należy ocenić bardzo pozytywnie w kontekście wyników dwóch poprzednich eksperymentów. Nawet przy 10% rozmyciu skupień *WKO* dla wybranych par zachowuje zdolność dyskryminacyjną, gdyż wartości *WKO* są w rozróżnialny sposób wyższe od wartości dla par zmiennych, z których jedna (lub obie) nie tworzą żadnej struktury skupień.

Wyniki powyższych eksperymentów pozwalają zaproponować znacznie prostszą metodę wybierania zmiennych binarnych tworzących strukturę skupień niż w przypadku zmiennych ciągłych lub porządkowych. W przypadku zmiennych binarnych trudniej jest badać zmiany wartości *WKO* przy różnego rodzaju wymuszeniach lub ograniczeniach zbioru obiektów, gdyż, na przykład, odrzucenie jednego z wariantów cechy uniemożliwia obliczanie *WKO*. Bezpiecznym progiem dla wartości *WKO*(*u,v*) , od którego możemy uznać, że pary zmiennych (*u, v*) są istotne dla tej samej struktury skupień może być liczba **0,05**. Jeśli założymy, że próg ten (uzyskany dla wektorów zmiennych binarnych o ustalonej długości i ustalonej postaci) będzie dobry dla innych struktur skupień tworzonych przez zmienne binarne, to selekcja zmiennych może polegać na tym, by zaczynając od pary z najwyższą wartością *WKO* dołączać do niej wszystkie zmienne, które są związane z nią przez wartość *WKO* przekraczającą 0,05. Dokładniej, metoda bezwzględna wybierania zmiennych przedstawiona jest poniżej.

Metoda bezwzględna *KO_binarne* selekcji zmiennych binarnych

Krok 0. Połóż *i*=0 oraz zbiór wszystkich zmiennych na miejsce zbioru zmiennych pozostałych do analizy.

Krok 1. Połóż *i*=*i*+1. Znajdź dla wszystkich par zmiennych (*u,v*) (ze zbioru zmiennych pozostałych do analizy) wartość *WKO*(*u,v*).

Krok 2. Znajdź wśród tych liczb znalezionych w kroku 1 najwyższą wartość *WKO*(*u,v*) . Parę odpowiadającą tej wartości oznacz przez (u_0, v_0).

Krok 3. Jeśli $WKO(u_0, v_0) < 0{,}05$, to idź do kroku 5.

Krok 4. Jeśli $WKO(u_0, v_0) \geq 0{,}05$, to dołączaj sekwencyjnie pojedyncze zmienne *s* do pary (u_0, v_0) jeżeli *WKO* dla zmiennej *s* i którejkolwiek spośród zmiennych już dołączonych lub zmiennych z pary (u_0, v_0) ma wartość wyższą od 0,05. Dołączanie sekwencyjne kontynuuj dotąd gdy nie będzie wśród zmiennych pozostałych do analizy żadnej zmiennej spełniającej ten warunek. Zbiór zmiennych dołączonych do pary (u_0, v_0) oznacz przez A_i. Połóż zbiór pozostałych zmiennych na miejsce zbioru zmiennych pozostałych do analizy i idź do kroku 1, gdy liczebność tego zbioru jest co najmniej 3. Gdy liczebność jest mniejsza od 3 , to idź do kroku 5.

Krok 5. Uznaj wszystkie wyodrębnione zbiory A_i za zbiory tworzące struktury skupień.

Metody porównawczej (wykorzystującej założenie o istnieniu co najwyżej jednej struktury skupień) wybierania zmiennych odpowiadającej powyższej metodzie bezwzględnej nie można sformułować, jako że uznajemy, że przyjęta

wartość progowa pozwalająca na wyodrębnianie rozłącznych podzbiorów zmiennych zapewnia również to, że każdy z tych podzbiorów tworzy swoją strukturę skupień. Można jedynie wśród wyodrębnionych podzbiorów wyróżnić ten, który ma najsilniejszą strukturę skupień. Wówczas należy przyjąć następującą zasadę: spośród wszystkich wyodrębnionych zbiorów A_i , za zbiór zmiennych tworzących najsilniejszą strukturę skupień uznaj ten, który ma najwyższą średnią arytmetyczną $WKO(u,v)$ ze wszystkich par (u,v) zmiennych z tego zbioru.

Metoda taka powinna spisywać się dobrze dla struktur skupień o podobnych parametrach jak struktury zaczerpnięte z pracy Brusco (2004) (por. 1.4.3). Błędne rezultaty mogą pojawić się na przykład wtedy, gdy liczba zmiennych będzie znacznie wyższa od kilku i niektóre skupienia będą słabo rozróżnialne z powodu słabej rozróżnialności poszczególnych zmiennych opisujących poszczególne skupienia. W punkcie 4.5.3 przedstawiona jest ocena efektywności obu metod.

4.5. Ocena nowych metod przy pomocy eksperymentów symulacyjnych

4.5.1. Eksperyment dla zbiorów danych ze zmiennymi ciągłymi

W celu zachowania porównywalności wyników do innych eksperymentów załóżmy, że rozważamy zawężony problem selekcji zmiennych tzn. w każdym zbiorze danych istnieje tylko jeden zbiór zmiennych tworzący pewną strukturę skupień. Taką cechę mają zbiory zestawu PODST. W takim problemie należy zastosować porównawczą wersję metody *KO_ciągłe.* Niemniej jednak, w ostatnim kroku metody *KO_ciągłe*, w którym mamy do wyboru decyzję bezwzględną i porównawczą, zawsze najpierw będziemy próbować podjąć decyzję bezwzględną (bez stosowania oceny wizualnej). Jeżeli to okaże się niemożliwe, to podjęta zostanie decyzja porównawcza, tak jak to zostało sformułowane w kroku 8 metody.

Wyniki eksperymentu przedstawione są w tabelach 4.18-4.23. Asymptotyczna odzyskiwalność (*ARI*) była obliczana w postaci średniej arytmetycznej skorygowanego indeksu Randa pomiędzy grupowaniem otrzymanym przy pomocy metody *k*-średnich (*k* – prawidłowa liczba skupień wynikająca ze sposobu generowania zbioru, 100 powtórzeń z losowym wyborem punktów startowych) a przynależnością obserwacji do skupień określoną przez sposób generowania zbiorów. Losowy wybór punktów startowych powielany wielokrotnie jest rekomendowany jako najlepsza metoda wyboru punktów startowych dla metody *k*-średnich (*Steinley i Brusco* ,

2007(b)). Poniżej przedstawiona jest szczegółowa analiza wyników metody *KO_ciągłe*.

Tabela 4.18 Efektywność metody *KO_ciągłe* względem liczby zmiennych nieistotnych oraz rodzaju struktury skupień tworzonej przez zmienne istotne.

Typ zbioru danych	Liczba zmiennych nieistotnych	Pamięć	Precyzja	*ARI*
Brak korelacji wewnątrz klas	2	0,887	0,917	0,713
	4	0,890	0,908	0,705
	6	0,892	0,894	0,704
Korelacja wewnątrz klas	2	0,910	0,934	0,651
	4	0,910	0,925	0,650
	6	0,906	0,915	0,646
Średnio		**0,899**	**0,916**	**0,678**

Źródło: obliczenia własne.

Z liczb przedstawionych w tabelach 4.18-4.23 widać, że metoda *KO_ciągłe* na tych samych zbiorach danych, na których oceniane były inne metody osiągnęła o lepsze wyniki od trzech najlepszych spośród pozostałych metod. Przede wszystkim zwraca uwagę stabilność metody w odniesieniu do liczby zmiennych nieistotnych dla struktury skupień i do skorelowania zmiennych wewnątrz skupień. Takiej cechy nie miała metoda HINoV i jej modyfikacje.

Tabela 4.19 Efektywność metody *KO_ciągłe* 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	0,847	0,928	0,875	0,878	0,889	0,936	0,941
Precyzja	0,861	0,945	0,895	0,894	0,896	0,953	0,963
ARI	0,649	0,685	0,677	0,682	0,677	0,687	0,688

Źródło: obliczenia własne.

Względem rodzaju zmiennych nieistotnych nie ma już takiej stabilności ale jest o wiele większa niż u konkurencji (por. tab. 4.19). Wyniki są ciekawe w kontekście skorelowania zmiennych wewnątrz skupień, metoda *KO_ciągłe* w sensie asymptotycznej odzyskiwalności spisała się nieco lepiej dla zbiorów bez korelacji wewnątrzklasowej natomiast w sensie pamięci i precyzji odwrotnie. Wydaje się, że efekt ten można wytłumaczyć tym, że zbiory z dodatnią korelacja wewnątrzklasową trudniej poprawnie pogrupować w skupienia, gdyż „wydłużone" skupienia są przeszkodą w odróżnianiu skupień dla zastosowanej metody grupowania tj. *k*-średnich. Natomiast zmienne w tym przypadku łatwiej dzieli się na różne podzbiory, gdyż występowanie korelacji wewnątrzklasowej skutkuje wyższymi wartościami *WKO* i *WKL*, co daje nieco lepsze rezultaty podziału zbioru zmiennych.

Tabela 4.20 Częstość rodzajów decyzji metody *KO_ciągłe*.

Rodzaj decyzji		Zbiory bez korelacji wewnątrzklasowej	Zbiory z korelacją wewnątrzklasową	Ogółem
Decyzja bezwzględna	Porównanie *WKO* z *WKL*	0,154	0,172	0,329
	Miara 4.6 >0,4	0,183	0,143	0,330
Decyzja względna		0,158	0,180	0,341

Źródło: obliczenia własne.

Z liczb zawartych w tabeli 4.20 wynika, że metoda *KO_ciągłe* dość często (prawie 33% decyzji) wykorzystywała porównanie *WKO* z *WKL* do podjęcia decyzji bezwzględnej. Równie często wysoka wartość miary (4.6) była podstawą decyzji bezwzględnej. W uzupełnieniu tabeli 4.20 należy dodać, że w decyzjom bezwzględnym towarzyszyła niemalże doskonała pamięć i precyzja – odpowiednio 0,994 i 0,999. Ten rezultat należy ocenić jako bardzo dobry. Z dotychczas opracowanych metod tylko podejścia modelowe podejmują decyzje bezwzględne ale efektywność tych metod jest słaba.

Tabela 4.21 przedstawia efektywność metody *KO_ciągłe* w zależności od liczby skupień w zbiorze danych. Wyniki są poprawne, gdyż efektywność rośnie wraz ze wzrostem liczby skupień. Ten wynik pozostaje w zgodności z badaniami wstępnymi (por. § 1.4) własności korelacji odległościowej, z których wynika, że gdy liczba skupień jest większa od 2, to zdolność dyskryminacyjna korelacji odległościowej jest większa. Uwagę zwraca znacznie słabsza od pozostałych wyników, efektywność w przypadku najmniejszej liczby skupień tj. dwóch skupień. Ten wynik również zgadza się z badaniami wstępnymi. Ponadto,

dla najwyższej liczby skupień (8) efektywność jest nieco mniejszych liczb skupień. To jest zapewne konsekwencją zastosowanego tylko dwustopniowego grupowania obiektów.

Tabela 4.21 Efektywność metody *KO_ciągłe* w zależności od liczby skupień w zbiorze danych.

Liczba skupień	Pamięć	Precyzja	*ARI*
2	0,712	0,730	0,322
3	0,902	0,936	0,648
4	0,954	0,966	0,783
6	0,966	0,972	0,830
8	0,962	0,972	0,807

Źródło: obliczenia własne.

Tabela 4.22 przedstawia asymptotyczną odzyskiwalność metody *KO_ciągłe* w zależności od liczby zmiennych tworzących strukturę skupień i nieistotnych dla niej. Można zauważyć, że w każdej kolumnie efektywność metody jest niemalże identyczna. Takiej stabilności nie miały HINoV i jej modyfikacje – efektywność malała wyraźnie wraz ze wzrostem liczby zmiennych nieistotnych.

Tabela 4.23 przedstawia efektywność metody *KO_ciągłe* w zależności od stopnia separowalności skupień. Można zauważyć bardzo regularne zmniejszanie się efektywności wraz ze wzrostem stopnia rozmycia skupień. Wyniki ten jest poprawny i zgodny z oczekiwaniami.

Tabela 4.22 Asymptotyczną odzyskiwalność metody *KO_ciągłe* w zależności od liczby zmiennych istotnych (wiersze) oraz liczby zmiennych nieistotnych (kolumny).

	2	4	6
2	0,641	0,632	0,633
4	0,686	0,686	0,685
6	0,718	0,715	0,708

Źródło: obliczenia własne.

Tabela 4.23 Efektywność metody *KO_ciągłe* w zależności od stopnia zachodzenia skupień na siebie (*overlap*).

Overlap	Pamięć	Precyzja	*ARI*
0	0,996	0,994	0,983
0,1	0,931	0,949	0,727
0,2	0,882	0,902	0,624
0,3	0,850	0,870	0,553
0,4	0,836	0,861	0,503

Źródło: obliczenia własne.

4.5.2. Eksperyment dla zbiorów danych ze zmiennymi porządkowymi

Przeprowadzono eksperyment badawczy w celu sprawdzenia efektywności metody dla zbiorów danych ze zmiennymi porządkowymi. W celu zachowania możliwości porównania wyników zastosowano wersję porównawczą metody tj. działającą przy założeniu, że istnieje tylko jedna struktura skupień. Zestaw zbiorów składał się z 400 (por. 1.4.2). Metoda porównawcza *KO_porządkowe* spisała się bardzo dobrze, w przypadku 7 spośród 8 modeli (czyli dla 350 spośród 400 zbiorów danych) metoda bezbłędnie wybrała wszystkie zmienne istotne dla struktury skupień. Dokładniejsze wyniki przedstawione są poniżej.

Model 1. Wybór pary najistotniejszej dla struktury skupień 7 razy w oparciu o *WKL* , 43 razy w oparciu o zmianę *WKL* (czyli $R(u,v)$). Podział na dwa zbiory zmiennych - bezbłędny. Decyzja o wyborze zbioru zmiennych istotnych - bezbłędna.

Model 2. Wybór pary najistotniejszej dla struktury skupień 50 razy w oparciu o *WKL*. Podział na dwa zbiory zmiennych - bezbłędny. Decyzja o wyborze zbioru zmiennych istotnych - bezbłędna.

Model 3. Wybór pary najistotniejszej dla struktury skupień 50 razy w oparciu o *WKL*. Podział na dwa zbiory zmiennych - bezbłędny. Decyzja o wyborze zbioru zmiennych istotnych - bezbłędna.

Model 4. Wybór pary najistotniejszej dla struktury skupień 50 razy w oparciu o *WKL*. Podział na dwa zbiory zmiennych - bezbłędny. Decyzja o wyborze zbioru zmiennych istotnych - bezbłędna.

Model 5. Wybór pary najistotniejszej dla struktury skupień 50 razy w oparciu o *WKL*. Podział na dwa zbiory zmiennych - bezbłędny. Decyzja o wyborze zbioru zmiennych istotnych - bezbłędna.

Model 6. Brak pary najistotniejszej dla struktury skupień, która miałaby $|WKL| > 0{,}3$ lub zmianę *WKL* większą od 0,3.

Model 7. Wybór pary najistotniejszej dla struktury skupień 5 razy w oparciu o *WKL* , 45 razy w oparciu o zmianę *WKL*. Podział na dwa zbiory zmiennych - bezbłędny. Decyzja o wyborze zbioru zmiennych istotnych - bezbłędna.

Model 8. Wybór pary najistotniejszej dla struktury skupień 50 razy w oparciu o zmianę *WKL*. Podział na dwa zbiory zmiennych - bezbłędny. Decyzja o wyborze zbioru zmiennych istotnych - bezbłędna.

W uzupełnieniu tych wyników należy zaznaczyć, że zastosowano metodę porównawczą ale decyzja o wyborze zbioru zmiennych istotnych zawsze była bezwzględna, za każdym razem zmiana wartości *WKL* była wyższa od 0,3.

4.5.3. Eksperyment dla zbiorów danych ze zmiennymi binarnymi

W celu ocenienia metody *KO_binarne* dla zbiorów danych ze zmiennymi binarnymi zastosowano obie metody tj. bezwzględną i porównawczą do zbiorów danych binarnych (por. 1.4.3), z pojedynczymi strukturami skupień.

Metoda porównawcza wybrała bezbłędnie zmienne w przypadku 710 zbiorów spośród 729 zbiorów. Ten wynik jest odrobinę gorszy od wyniku metody Brusco, ale nie zostało wykorzystane założenie o znanej liczbie skupień.

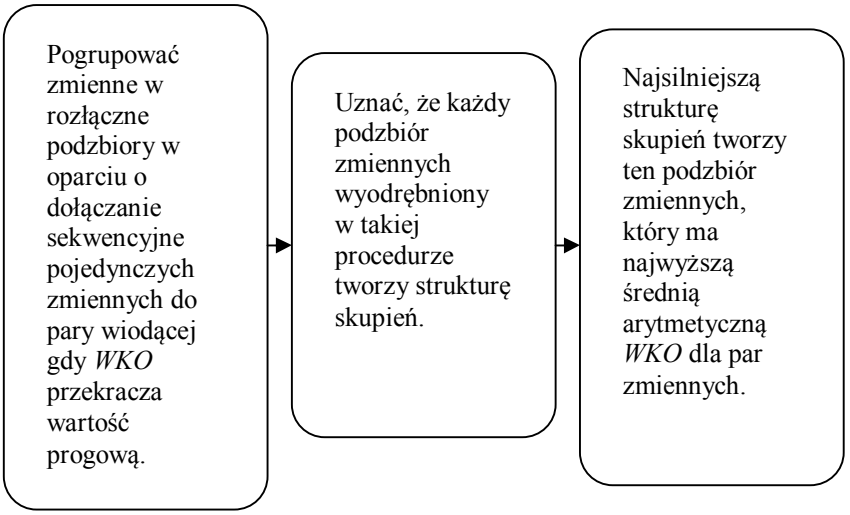
Metoda bezwzględna spisała się w identyczny sposób pod względem liczby zbiorów z bezbłędnie wybranymi zmiennymi. W przypadku pozostałych 19 zbiorów, w 7 przypadkach błędnie wyróżniła drugi podzbiór zmiennych (zawsze była to para zmiennych), jako zbiór tworzący drugą strukturę skupień.

Metodę bezwzględną zastosowano również do trzech zbiorów z wielokrotnymi strukturami skupień (por. 1.4.3). Metoda w każdym zbiorze wybrała bezbłędnie wszystkie 12 zmiennych (2 razy po 6 zmiennych) opisujących dwie niezależne od siebie struktury skupień. Pozostałe 6 zmiennych odrzuciła jako zmienne nieistotne dla żadnej struktury skupień. Wynik ten należy ocenić bardzo pozytywnie, zwłaszcza w przypadku trzeciego zbioru z bardzo dużym (5%) rozmyciem obu struktur skupień. Wynik ten jest nieporównywalny z innymi metodami, gdyż nie są znane metody, które potrafiłyby wybierać binarne zmienne istotne dla wielokrotnych struktur skupień.

Przeprowadzone badanie potwierdza uniwersalne możliwości korelacji odległościowej.

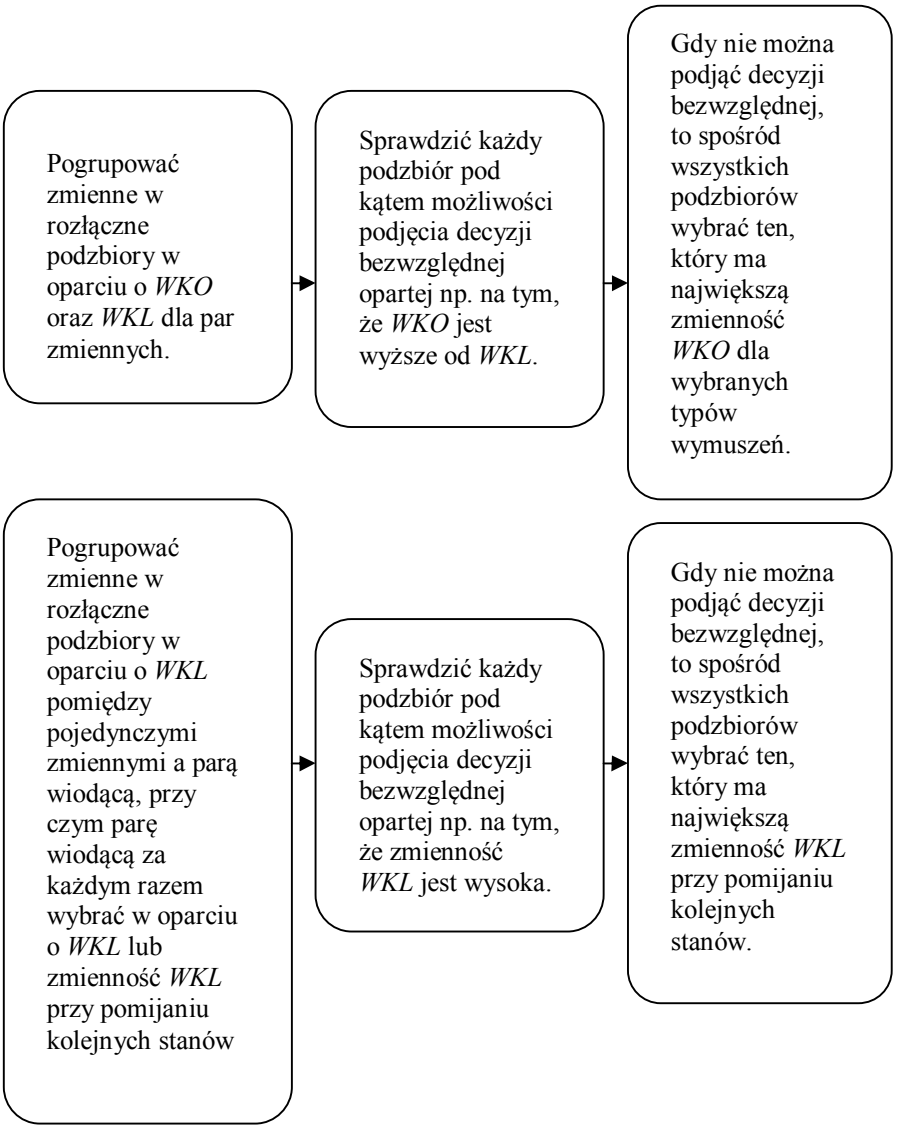
4.6. Sformułowanie metody dla dowolnego układu skal pomiarowych

Na rysunkach 4.6 i 4.7 przedstawiona jest idea działania zaproponowanych metod selekcji dla zmiennych mierzonych na skalach o jednakowej sile pomiaru każdej zmiennej. W każdym algorytmie nacisk położony jest na jak najbardziej optymalne wykorzystanie właściwości danej skali pomiarowej. Wśród tych metod nie ma algorytmu dla zmiennych nominalnych wielostanowych. Ponadto, konieczne wydaje się zaproponowanie metody ogólnej tj. dla dowolnego zestawu rodzajów skal pomiarowych z powodu bardzo często spotykanych w praktyce zbiorów tego typu przy towarzyszącym temu kompletnym niedostatku metod selekcji zmiennych w takich przypadkach.



Rys. 4.6 Diagram przedstawiający zasadę działania metody selekcji zmiennych binarnych.
Źródło: opracowanie własne.

Najbardziej naturalnym podejściem do zagadnienia konstrukcji algorytmu selekcji zmiennych wtedy gdy mamy do czynienia z występowaniem mieszanych skal pomiarowych jest oparcie go na zasadzie podobnej do reguły sekwencyjnego dołączania zmiennych do jakiejś pary wyjściowej/wiodącej występującej w każdym z algorytmów z rysunku 4.6. Występowanie skal pomiarowych o różnej sile było, w dotychczasowym dorobku nauki w temacie



Rys. 4.7 Diagram przedstawiający zasady działania dwóch różnych metod selekcji zmiennych w zależności od rodzaju skal pomiarowych. U góry metoda dla zmiennych ciągłych, u dołu metoda dla zmiennych porządkowych.

Źródło: opracowanie własne.

selekcji zmiennych w analizie skupień, na ogół, źródłem problemów. Nieliczne są metody, które dopuszczają występowanie różnych skal pomiarowych w jednym zbiorze danych i spisują się one bardzo słabo. Można jednak zaproponować metodę, w której wykorzystamy niewątpliwą zaletę słabszych skal pomiarowych, tę, którą wykorzystali Fisher i Talavera (por. § 3.3). Zaletą tą

jest to, że fakt wysokiego skorelowania dwóch zbiorów zmiennych mierzonych na skalach dyskretnych (o małej liczbie stanów) świadczy o istnieniu struktury skupień dla zbioru zmiennych, który jest sumą mnogościową obu zbiorów. Problemem jest pomiar siły korelacji, ale do tego zadania można wykorzystać korelację odległościową, która jest miarą uniwersalną w tym sensie, że można ją stosować do mierzenia siły wzajemnego wkładu dwóch zbiorów zmiennych do tworzenia tej samej struktury skupień.

Algorytm, który zostanie zaproponowany będzie działał na podobnej zasadzie i, przy jego sformułowaniu, będą wykorzystane własności korelacji odległościowej oraz liniowej, które można było zaobserwować wśród przykładowych struktur skupień zbadanych przy konstruowaniu algorytmów *KO_porządkowe* oraz *KO_binarne.* W szczególności, można zauważyć, że wraz ze spadkiem dokładności pomiaru spadają wymagania odnośnie minimalnej wartości *WKO* jaka może świadczyć o istnieniu struktury skupień. Dla zmiennych ciągłych mierzonych na skali ilorazowej progiem takim była wartość 0,15. Dla przykładowych struktur zmiennych binarnych (por. tab. 1.4) wartość 0,1 świadczy w sposób zdecydowany o istnieniu struktury skupień. Ponadto należy wziąć pod uwagę własności *WKO* odnoszące się do zależności od liczebności zbiorów zmiennych pomiędzy którymi ten współczynnik jest obliczany. Im większe liczebności zbiorów zmiennych tym łagodniejsze wymagania dla wartości *WKO*, która świadczy o istnieniu struktury skupień. W związku z tym próg powyżej którego będzie można przyłączać pary zmiennych do innego (o liczebności co najmniej 2) zbioru zmiennych zostanie obniżony do 0,1. Algorytm, który nazwiemy *KO_meta* będzie zatem działał na zasadzie sekwencyjnego dołączania zmiennych (lub par zmiennych), do pary wyjściowej, którą może być dowolna para zmiennych, dla której *WKO* przekracza minimalny próg (0,15). Jeżeli zmiennych będzie kilka (mniej niż 10), to będą dołączane pojedyncze zmienne. Gdy zmiennych będzie więcej, to będą dołączane pary zmiennych. Po wyodrębnieniu rozłącznych podzbiorów zmiennych będzie można wykorzystać wspomnianą zaletę zmiennych mierzonych na słabych skalach do podjęcia ewentualnej decyzji bezwzględnej o tym, czy każdy z wyodrębnionych podzbiorów zmiennych można uznać za zbiór tworzący strukturę skupień. Jeśli w podzbiorze *A* zmiennych, spośród wyodrębnionych podzbiorów zmiennych znajdziemy dwa rozłączne podzbiory, z których co najmniej jeden jest podzbiorem zmiennych mierzonych na kilkustanowych skalach dyskretnych, dla których *WKO* przekracza minimalny próg, to podzbiór *A* uznamy za zbiór tworzący strukturę skupień. Jeżeli decyzji bezwzględnej nie będzie można podjąć dla żadnego z wyodrębnionych podzbiorów, to podjęta zostanie decyzja o charakterze porównawczym, w której kryterium będzie podobne do kryterium z algorytmu *KO_ciągłe.* Poniżej przedstawione jest dokładne sformułowanie algorytmu *KO_meta.*

Metoda *KO_meta* selekcji zmiennych tworzących strukturę skupień

Krok 1. Znajdź dla wszystkich par zmiennych (u,v) (ze zbioru zmiennych pozostałych do analizy) wartość $WKO(u,v)$.

Krok 2. Znajdź wśród liczb znalezionych w kroku 1 najwyższą wartość $WKO(u,v)$. Parę odpowiadającą tej wartości oznacz przez (u_0, v_0).

Krok 3. Jeśli $WKO(u_0, v_0) < 0{,}15$, to zakończ postępowanie podejmując decyzję o braku zmiennych tworzących strukturę skupień gdy para znaleziona w kroku 2 była tylko jedna lub idź do kroku 8 gdy par tych było więcej.

Krok 4. Jeśli $WKO(u_0, v_0) \geq 0{,}15$, to oznacz $A = \{u_0, v_0\}$.

Krok 5. Wśród wszystkich par (u,v) pozostałych zmiennych znajdź parę (u_1, v_1), dla której $WKO\{A,(u_1, v_1)\}$ ma najwyższą wartość.

Krok 6. Jeżeli $WKO(A,(u_1, v_1)) > 0{,}10$, to połóż $A = A \cup \{u_1, v_1\}$. Idź do kroku 5. Jeżeli $WKO(A,(u_1, v_1)) \leq 0{,}10$, to zakończ postępowanie uznając zbiór A zmiennych za zbiór wyodrębniony.

Krok 7. Powtórz kroki 2-6 dla pary startowej w kroku 2, która ma najwyższą wartość *WKO* spośród wszystkich par, które można utworzyć ze zmiennych pozostałych do analizy. Jeżeli nie ma pary spełniającej warunek $WKO(u,v) \geq 0{,}15$, to zakończ wyodrębnianie podzbiorów zmiennych i przejdź do kroku 8.

Krok 8. Sprawdź każdy z wyodrębnionych podzbiorów pod kątem poprawności dołączonych par zmiennych przy $0{,}1 < WKO(A,(u_1, v_1)) < 0{,}2$ (patrz Uwaga 1 poniżej). Sprawdź każdy z wyodrębnionych podzbiorów pod kątem możliwości dołączenia pojedynczych zmiennych, które w innym ciągu dołączeń były dołączane do zbioru zmiennych będącego podzbiorem wyodrębnionego podzbioru (patrz Uwaga 2 poniżej). Oceń każdy z wyodrębnionych podzbiorów zmiennych przy pomocy następujących kryteriów bezwzględnych:

- wartość miary (4.6) większa od 0,4 w przypadku gdy wyodrębniony podzbiór zmiennych składa sie ze zmiennych tylko ciągłych;
- występowanie jakichkolwiek dwóch podzbiorów (w szczególności dwóch par zmiennych oraz pary zmiennych z jedną zmienną) *B, C* wyodrębnionego podzbioru takich, że $WKO(B,C) \geq 0{,}1$ oraz co najmniej jeden z podzbiorów *B, C* składa się tyko ze zmiennych nominalnych lub porządkowych.

Jeżeli żaden spośród wyodrębnionych podzbiorów nie spełnia żadnego z kryteriów, to idź do kroku 9. Jeżeli istnieją jakieś podzbiory spełniające jedno z

kryteriów bezwzględnych, to uznaj je za zbiory tworzący strukturę skupień i zakończ postępowanie.

Krok 9. Jeśli wśród wyodrębnionych podzbiorów zmiennych są zbiory zmiennych tylko ciągłych, to wybierz jako zbiór zmiennych tworzących strukturę skupień, zbiór, który ma najwyższą wartość miary (4.6).

Kryterium porównawcze z kroku 9 powinno być skonstruowane tak, by można było porównać dwa podzbiory zmiennych składające się ze zmiennych, które mogą być mierzone na różnych skalach. Takie kryterium musiałoby jednak być bardzo rozbudowane i oparte na sztucznych progach, których wysokości byłyby trudne do uzasadnienia. Zauważmy, że decyzja porównawcza, w przypadku takich podzbiorów, nie ma takiego znaczenia jak, na przykład, w przypadku porównywania dwóch podzbiorów zmiennych ciągłych. Tak jest z tego względu, że jeśli zmienne mierzone na różnych skalach tworzą strukturę skupień, to musi to znaleźć odzwierciedlenie w wysokich wartościach *WKO* w ramach każdego podzbioru. Innymi słowy, jeżeli zmienne mierzone na różnych skalach tworzą strukturę skupień, to w kroku 8 powinniśmy tę własność ustalić przy pomocy jednego z kryteriów bezwzględnych. Jeśli takiej decyzji nie podejmiemy, to zasadniczo rzecz biorąc, należy podjąć decyzję o braku struktury skupień w tym podzbiorze. A zatem kryterium porównawcze z kroku 9 będzie miało marginalne znaczenie – będzie stosowane tylko do porównania dwóch podzbiorów zmiennych składających się tylko ze zmiennych ciągłych, co pozwala takie sformułowanie kryterium.

Uwaga 1. Jak zaznaczono wcześniej, w przypadku małej liczebności zbioru wszystkich zmiennych należy do pary wyjściowej dołączać raczej zmienne pojedyncze. Różnica w sformułowaniu algorytmu będzie więc polegała na tym, że w kroku 5, do zbioru *A* dołączamy zmienną *u* dającą najwyższą wartość $WKO(A,u)$. Gdy łączenie zmiennych odbywa się na parach zmiennych, to może zdarzyć się, że zmienna nieistotna dla struktury zostanie włączona do zbioru zmiennych istotnych w ramach pary ze zmienną istotną, bardzo silnie skorelowaną z innymi zmiennymi istotnymi. Dołączenie takiej pary powinno odbyć się przy niezbyt wysokiej (bo jedna zmienna tworzy strukturę skupień a druga nie) wartości *WKO,* np. $0{,}1 < WKO(A,(u_1,v_1)) < 0{,}2$. Wobec tego przynależność każdej zmiennej *u* dołączonej do danego zbioru wyodrębnionego przy wartości *WKO* mniejszej od 0,2 należy sprawdzić. Tego sprawdzenia można dokonać znajdując wartości $WKO(A,u)$ dla każdej ze zmiennych *u,* ze wszystkimi podzbiorami *A* zbioru wyodrębnionego o liczebności co najmniej 2. Jeżeli nie znajdziemy podzbioru o liczebności co najmniej 2, z którym zmienna *u* miałaby $WKO(A,u) \geq 0{,}1$, ani podzbioru jednoelementowego *A*, z którym

zmienna u miałaby $WKO(A,u) \geq 0{,}15$, to zmienną u wykluczamy ze zbioru wyodrębnionego.

Uwaga 2. Rozbieżności dotyczące dołączania lub wykluczania zmiennych ze zbiorów wyodrębnionych mogą pojawić się również w sytuacji gdy pojedyncze zmienne nie są włączone do danego zbioru pomimo, że w innym ciągu dołączeń były dołączane do zbioru składającego się, w większości, z tych samych zmiennych. Wobec tego zaproponujmy następującą zasadę: jeżeli zmienna *u,* nienależąca do wyodrębnionego podzbioru zmiennych *A,* była w innym ciągu dołączeń dołączona do innego zbioru, który zawierał co najmniej połowę zmiennych składających się na zbiór *A*, to zmienną *u* należy dołączyć do zbioru *A*.

Zaproponowany algorytm nie będzie badany w eksperymencie symulacyjnym z powodu zbyt daleko idącej dowolności w konstruowaniu struktur skupień zmiennych mierzonych na różnych skalach. W literaturze nie spotyka się eksperymentów symulacyjnych dla zbiorów danych opisanych, na przykład, tylko zmiennymi nominalnymi. Metody analizy skupień (nie tylko selekcji zmiennych) poświęcone takim zbiorom badane są na przykładowych zbiorach empirycznych. W rozdziale 5 efektywność algorytmu *KO_meta* zostanie oceniona w oparciu o zastosowanie do dużej liczby zbiorów empirycznych.

5. ZASTOSOWANIE METOD SELEKCJI ZMIENNYCH W ANALIZIE SKUPIEŃ W BADANIACH EKONOMICZNO-SPOŁECZNYCH

5.1. Zagadnienia wstępne dotyczące empirycznych zbiorów danych

Dobra efektywność metod selekcji zmiennych w eksperymencie na zbiorach ze zmiennymi ciągłymi, w postaci mieszanin rozkładów normalnych oraz w eksperymentach na zbiorach ze zmiennymi mierzonymi na słabszych skalach, generowanymi w oparciu o założoną postać struktury skupień jest w pewnym sensie warunkiem koniecznym dobrej całościowej oceny metody. W analizie skupień ocena ostateczna powinna być jednak oparta na efektywności metody w zastosowaniu do empirycznych zbiorów danych. Empiryczne zbiory danych często cechuje występowanie różnych skal pomiarowych oraz skupień, dla których modelem (dla silnych skal) nie zawsze może być rozkład normalny. W przypadku występowania zmiennych z różnymi skalami pomiarowymi jedyną zaproponowaną dotychczas metodą może być metoda Ichino oparta na teorii grafów. Ta metoda ma jednak poważne ograniczenia natury obliczeniowej, oraz jest bardzo nieefektywna. Ocena oparta na empirycznych zbiorach danych powinna być przeprowadzona w oparciu o zbiory sprawdzone przez specjalistów z danej dziedziny. Takie zbiory można znaleźć w literaturze oraz wielu bazach danych, z których najbardziej popularna jest baza UCI czyli repozytorium danych Uniwersytetu Kalifornijskiego.

Bardzo istotnymi zagadnieniami, do których należy się ustosunkować są:

- dobór metody normalizacji zmiennych;
- dobór miary odległości dla różnych skal pomiarowych.

Te dwie kwestie są ze sobą związane (jak zaznaczono w § 1.2), ponieważ niekiedy formuła normalizacyjna ma wpływ na wybór miary odległości (por. *Gatnar, Walesiak*, 2004). Do pomiaru odległości pomiędzy obserwacjami empirycznych zbiorów danych zastosowane zostaną te same miary, które stosowane były w przypadku zbiorów danych występujących w eksperymentach w dotychczasowej części pracy (por. tabela 1.1). W przypadku występowania różnych skal pomiarowych zastosowana zostanie odległość dana wzorem (1.10), która budzi najmniej zastrzeżeń spośród wszystkich propozycji ujednolicania odległości uzyskanych z różnych skal pomiarowych. W przypadku zmiennych mierzonych na silnych skalach pomiarowych (interwałowej i ilorazowej)

zastosowana zostanie również formuła normalizacyjna taka jak w dotychczasowej części pracy tj. formuła (1.2).

Istotną kwestią jest również dobór metod konkurencyjnych, w porównaniu z którymi będziemy oceniać zaproponowane metody. W świetle ocen podsumowujących z rozdziału 2 i 3 wybór wydaje się raczej zdeterminowany. Metody modelowe wybierania zmiennych okazały się dużo słabsze od heurystycznych w eksperymencie, w którym zbiory były generowane z mieszanin rozkładów normalnych (por. *Steinley i Brusco*, 2008). Na tej podstawie, biorąc pod uwagę również to, że zbiory ze świata realnego rzadko da się wymodelować w tej postaci (a taki model zbioru danych zakładają metody modelowe – por. wzór (2.1)), pomijamy szukanie konkurencji wśród podejść modelowych. Na przykład, w repozytorium zbiorów UCI trudno znaleźć zbiór, który miałby wszystkie skupienia o rozkładzie normalnym. W zbiorze *Donation* ani pierwsze, ani drugie skupienie nie mają rozkładu normalnego na żadnej z czterech zmiennych, w zbiorze *Concrete* drugie skupienie (91 obiektów spośród 1030) nie ma normalności na większości z ośmiu zmiennych, w zbiorze *Glass* pierwsze skupienie (70 obiektów spośród 214) nie ma normalności na większości z dziewięciu zmiennych, w zbiorze *Iris* pierwsze skupienie (50 obiektów spośród 150) nie ma normalności na czwartej zmiennej, w zbiorze *Wines* pierwsze skupienie (59 obiektów spośród 178) nie ma normalności na drugiej zmiennej zaś drugie skupienie (71 obiektów) nie ma normalności ani na drugiej, ani na piątej zmiennej (test Kołmogorowa-Smirnowa, obliczenia w pakiecie SPSS 19). Ciekawe jest na przykład to, że zbiór *Wines* był użyty jako zbiór przykładowy do zilustrowania działania podejścia wyboru modelu w pracy Raftery i Dean (2006). Fakt ten świadczy o tym, że trudno było znaleźć zbiór, który spełniałby wszystkie założenia potrzebne do otrzymania wyników w postaci wnioskowania statystycznego.

Spośród metod heurystycznych wybrane zostały trzy metody, które można uznać za najlepsze spośród dotychczas opracowanych. A zatem, będziemy zatem oceniać zaproponowane metody w porównaniu z metodą HINoV. Jeśli będzie to możliwe, to będziemy również uwzględniać obie modyfikacje tej metody (VS-KM oraz VAF z indeksem skupialności). Będzie brana pod uwagę postać tych metod z kryterium liczbowym wyboru zmiennych (kryterium największego skoku por. § 3.6) jak również postacie oryginalne, opracowane dla pojedynczych zbiorów danych, polegające w przypadku dwóch z tych metod, na wizualnym badaniu wykresu osypiska. W przypadku większości zbiorów badanych w tym rozdziale nie ma efektywnej metody, z którą można by było porównać metodę *KO_meta* wobec tego przedstawiona będzie tylko efektywność naszej propozycji.

5.2. Metodologia oceny dla empirycznych zbiorów danych

W literaturze można spotkać się z dwoma sposobami oceny efektywności metod wybierania zmiennych w analizie skupień.

Pierwsza to opinie ekspertów z danych dziedzin, którzy wskazują zmienne ich zdaniem istotniejsze od innych zmiennych dla struktury skupień. Typowym przykładem może być zbiór *Iris* z bazy UCI Uniwersytetu Kalifornijskiego, w którym trzecia i czwarta zmienna są, zdaniem specjalistów, istotniejsze od dwóch pierwszych (por. np. *Dash i Liu*, 2000). Ta metoda nie zostanie zastosowana w niniejszej rozprawie, bo ma zbyt poważne wady. Po pierwsze opinie ekspertów mogą być subiektywne, po drugie, skąd wziąć ekspertów do wszystkich zbiorów, które chcielibyśmy zbadać, po trzecie, gdy zmiennych jest więcej np., kilkadziesiąt, to żaden ekspert nie ogarnie struktury ich wartości pamięciowo – musiałby posługiwać się w pewnym stopniu jakimś algorytmem komputerowym, który pomógłby mu wyszukiwać zależności pomiędzy zmiennymi.

Druga metoda, to metoda, która była zastosowana w eksperymentach symulacyjnych z rozdziału 4 tj. dołączanie do zbioru zmiennych nowych zmiennych zanieczyszczających istniejącą strukturę skupień zbioru danych i sprawdzanie czy badana metoda wybierania zmiennych istotnych poprawnie wyszuka zmienne tworzące strukturę skupień. Te metodę zastosujemy tylko w formie dołączania zmiennych zanieczyszczających, nie będzie przeprowadzana selekcja zmiennych na zbiorze danych oryginalnych. Uzasadnieniem jest to, zaproponowany algorytm *KO_meta*, daje wynik na ogół niezależny od tego czy dołączone zostały jakieś zmienne nieistotne – problemem jest poprawny wybór podzbioru zmiennych spośród wszystkich wyodrębnionych podzbiorów. Metoda dołączania zmiennych nieistotnych nie jest pozbawiona wad. Po pierwsze, można dołączać tylko rozkłady konkretnego typu, które nie muszą być dobrymi modelami rozkładów zmiennych ze świata realnego nie mających znaczenia dla struktury skupień. Po drugie, istnieją pewne ograniczenia stosowalności tej metody. Najszerszy wachlarz rozkładów zanieczyszczających można zastosować w przypadku zmiennych ciągłych. W przypadku zmiennych mierzonych na słabszych skalach, praktycznie rzecz biorąc, można dołączać tylko zmienne o równomiernym rozłożeniu wszystkich swoich wariantów na zbiorze wszystkich obiektów zbioru. Bardziej złożone schematy dołączania zmiennych zakłócających w przypadku słabych skal pomiarowych łączą się z zagadnieniem występowania wielokrotnych struktur skupień i skorelowaniem zmiennych (por. § 1.3), co ma bezpośrednie przełożenie na uwarunkowanie procesu selekcji zmiennych. Pomimo wspomnianych wad, tę metodę zastosujemy odpowiednio dobierając rozkłady zmiennych dołączanych, zakłócających strukturę skupień, gdyż wydaje się, że metoda ta jest najlepsza. Dobieranie zmiennych

zakłócających będzie polegało najczęściej na ustaleniu liczby poszczególnych rodzajów zmiennych, gdyż, na przykład, rodzaj skal pomiarowych nie powinien decydować o tym czy zmienne o danym typie skali można dołączać do danego zbioru zmiennych czy też nie. Stanowisko takie można uzasadnić tym, że trudno sobie wyobrazić jakiś racjonalny powód po temu by w początkowym etapie eliminacji zmiennych nieistotnych dla struktury skupień od razu wyeliminować, na przykład, zmienne ciągłe. Takie sytuacje mogą mieć miejsce w naukach o charakterze technicznym ale nie w naukach ekonomiczno-społecznych. Na przykład zbiór *Invest* składający się z funduszy powierniczych opisanych przez zmienne typowo ciągłe (mierzące ich efektywność) może być również opisany przez zmienną typowo nominalną jaką jest rodzaj funduszu (agresywne, zrównoważone itp.). Z kolei do zmiennych ze zbioru *Votes*, które wszystkie są nominalnymi binarnymi można dołączyć zmienną typowo ciągłą. Obiektami w tym zbiorze są kongresmeni, więc, na przykład, wielkość majątku mierzona w dolarach powinna być bardzo uzasadniona jako cecha wpływająca na to jak dany kongresmen głosował nad daną ustawą. W celu uniknięcia nadmiernej liczby rozpatrywanych wariantów dołączanych zbiorów zmiennych ustalmy liczbę wariantów nieprzekraczającą 2. W obu wariantach liczba dołączanych zmiennych zakłócających będzie mniej więcej taka sama jak liczba zmiennych oryginalnych. Natomiast zmienne będą różniły się między sobą innymi charakterystykami lub, w obu wariantach, charakterystyki zmiennych będą takie same, ale liczba zmiennych w jednym wariancie będzie mniej więcej równa liczbie zmiennych oryginalnych, zaś w drugim dwa razy większa. Na początku paragrafu 5.3 zbadanych zostanie kilka zbiorów opisanych tylko przez zmienne ciągłe w celu oceny porównawczej metody *KO_ciągłe*. Ta metoda, zbadana w poprzednim rozdziale, w obszernym eksperymencie symulacyjnym, jest ważniejsza od metod *KO_porządkowe* i *KO_binarne*, których skonstruowanie miało na celu pokazanie uniwersalnych możliwości korelacji odległościowej. W przypadku analizy empirycznych zbiorów danych, z cechami często mierzonymi na różnych skalach, znaczenie tych dwóch metod jest znikome. W dalszym ciągu paragrafu 5.3 badane będą zbiory danych, które po dodaniu zmiennych zakłócających będą miały cechy mierzone na różnych skalach. Analiza replikacji stosowana dla HINoV i jej modyfikacji będzie miała taka samą parametryzację jak w eksperymencie symulacyjnym tzn. 20 powtórzeń oraz grupowanie obiektów przy pomocy metody *k*-średnich z losowym wyborem obiektów startowych powtarzanym 50 razy. Gdy wszystkie zmienne (oryginalne i zanieczyszczające) będą ciągłe (mierzone na skali ilorazowej), to stosowana będzie odległość euklidesowa. W przypadku różnych skal pomiarowych stosowana będzie formuła (1.10) wraz z formułą (1.3) odległości GDM.

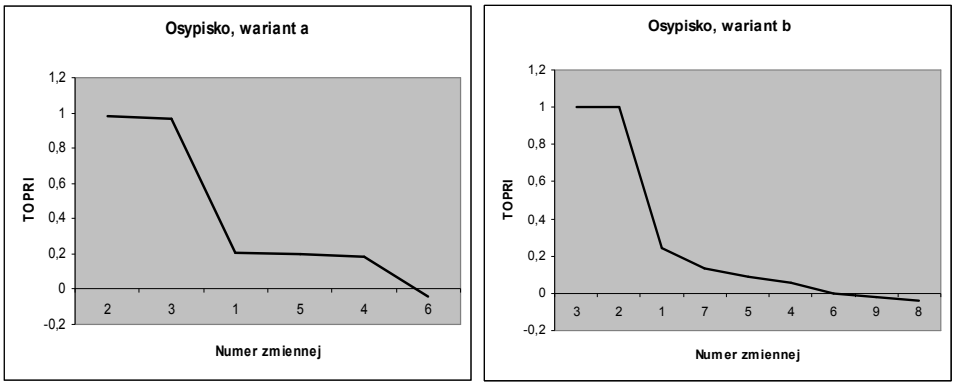
5.3. Badanie efektywności metod na empirycznych zbiorach danych

Zbiór ***Invest.*** Obiekty: fundusze inwestycyjne we Włoszech od 1996 roku. Liczebność: 103. Liczba zmiennych: 3. Liczba skupień: 2. Liczebności skupień: 56, 57.
Charakterystyka skupień: pierwsze skupienie obejmuje fundusze efektywniejsze, drugie skupienie fundusze mniej efektywne.
Charakterystyka zmiennych:

- pierwsza zmienna ciągła, efektywność funduszu z ostatnich 12 miesięcy;
- druga zmienna ciągła, efektywność funduszu z ostatnich 36 miesięcy;
- trzecia zmienna ciągła, zmienność wartości jednostki w ostatnich 36 miesiącach.

Zbiór zaczerpnięty z książki *Atkinson i inni,* (2004). Autorzy bardzo dokładnie zbadali ten zbiór dochodząc do wniosku, że zawiera on 2 lub 3 obserwacje nie pasujące do żadnego skupienia. Obserwacje te możemy uznać za nietypowe. Zbadajmy efektywność wybranych metod w dwóch wariantach dołączanych zbiorów zmiennych.

Wariant a) dwie zmienne (o numerach 4, 5) o dwuwymiarowym rozkładzie normalnym z zerowym wektorem średnim i macierzą kowariancji z jedynkami na głównej przekątnej i wartościami równymi 0,75 poza przekątną; trzecia zmienna (o numerze 6) o rozkładzie równomiernym na odcinku [0, 10]. Razem do analizy jest 6 zmiennych.



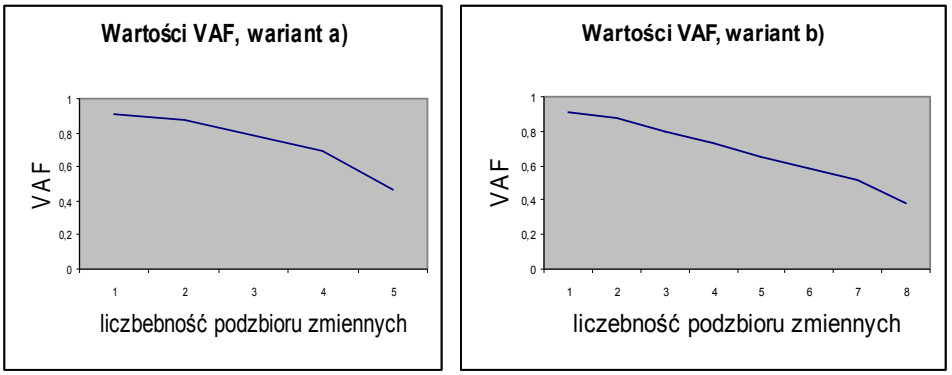
Rys. 5.1 Wykresy osypiska metody HINoV dla dwóch wariantów zbiorów zmiennych dla zbioru *Invest.*

Źródło: Opracowanie własne.

Wariant b) dwie zmienne (o numerach 4, 5) o dwuwymiarowym rozkładzie normalnym z zerowym wektorem średnim i macierzą kowariancji z jedynkami na głównej przekątnej i wartościami równymi 0,5 poza przekątną; trzecia zmienna (o numerze 6) o rozkładzie równomiernym na odcinku [0, 10]; czwarta zmienna (o numerze 7) o skośnym rozkładzie beta(1,1); piąta zmienna (o numerze 8) o rozkładzie normalnym ze średnią równą zero i odchyleniem standardowym równym 3; szósta zmienna (o numerze 9)) o rozkładzie normalnym ze średnią równą zero i odchyleniem standardowym równym 1. Razem do analizy jest 9 zmiennych.

Zarówno analiza replikacji jak i kryterium sylwetkowe zdecydowanie najczęściej wskazywały, że najlepszą liczbą skupień jest poprawna liczba 2 w obu wariantach a) oraz b) zbiorów zmiennych.

Metodę HINoV zastosowano dla liczby skupień równej 2. Metodą grupowania była metoda *k*-średnich z powtórzonym 50 razy losowym wyborem punktów startowych (por. § 3.6). Zarówno kryterium największego skoku metody HINoV jak i badanie wizualne wykresów osypiska (por. rys. 5.1) zwracają ten sam wynik. Metoda spisała się przeciętnie wybierając w obu wariantach dwie zmienne, drugą i trzecią pomijając zmienną pierwszą. Bardzo podobne wyniki uzyskano przy zastosowaniu zamiast grupowania metodą *k*-średnich aglomeracji metodą Warda lub średniego połączenia.



Rys. 5.2 Wykresy wartości VAF dla metody VAF z indeksem skupialności, w dwóch wariantach zbiorów zmiennych, dla zbioru *Invest*.

Źródło: Opracowanie własne.

Dla metody VSKM analizę przeprowadzono również dla poprawnej liczby 2 skupień – taka była zdecydowanie najczęściej wskazywana w analizie replikacji. W obu wariantach metoda spisała się bardzo słabo. W wariancie a) metoda wybrała tylko dwie zmienne: o numerach 2 i 5. W wariancie b) metoda również wybrała tylko dwie zmienne: o numerach 3 i 7 .

Dla metody VAF z indeksem skupialności analizę przeprowadzono również dla poprawnej liczby 2 skupień – taka była najczęściej wskazywana w analizie replikacji, w obu wariantach. W obu wariantach metoda spisała się bardzo źle. W obu wariantach, we wstępnym etapie selekcji, w oparciu o indeks skupialności odrzuciła błędnie pierwszą zmienną. Wartość indeksu, co prawda, nie była bardzo niska, 0,305 wobec progowej 0,3078 (por. § 3.8), ale była za niska. Wynik ten bardzo źle świadczy o efektywności indeksu, który w zamyśle autorów modyfikacji metody HINoV miał eliminować tylko zmienne nie przyczyniające się do generowania struktury skupień. W wariancie a) metoda wybrała 4 zmienne: drugą i trzecią (istotne) i czwartą i piątą (nieistotne). W wariancie b) metoda wybrała siedem zmiennych: od drugiej do ósmej (tylko druga i trzecia są istotne). Te wybory, od strony wizualnej, są przedstawione na rysunku 5.2. W przypadku a) spadek wartości VAF jest wyraźniejszy ale w obu wariantach wybrane zostały złe zmienne.

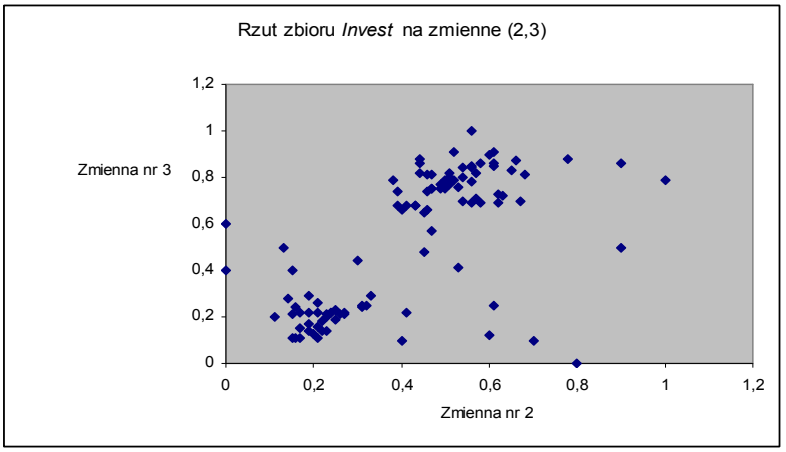
Tabela 5.1 Wartości *WKO, WKL* dla wybranych par zmiennych zbioru *Invest* w obu wariantach zmiennych zanieczyszczających.

Para zmiennych	*WKL*	*WKO*	*WKL*[7]	*WKO*[7]
{1, 2}	0,53	0,32	0,56	0,51
{1, 3}	0,29	0,07	0,49	0,25
{2, 3}	0,83	0,71	0,95	0,94
{4,5} war. a)	0,70	0,40	0,76	0,44
{4,5} war. b)	0,46	0,11	0,50	0,16

Źródło: Opracowanie własne.

Metoda *KO_ciągłe* w wariancie a) spisała się bezbłędnie. W tabeli 5.1 podane są wartości współczynników korelacji liniowej i odległościowej dla wybranych par zmiennych. Dla pozostałych par wszystkie wartości *WKL* oraz *WKO* były mniejsze od 0,1 co do wartości bezwzględnej. W pierwszym etapie, na bazie wartości *WKL* oraz *WKO,* metoda *KO_ciągłe* podzieliła zbiór wszystkich zmiennych na 3 podzbiory: {1,2,3}, {4,5} oraz {6}. Spośród tych trzech podzbiorów należy wybrać podzbiór {1,2,3} i decyzja ma charakter bezwzględny. Można to uzasadnić na kilka sposobów. Dla zbioru {1,2,3} wartość miary (4.6) równa jest 0,877 (znacznie wyższa od 0,4), takiej wartości nie uzyskałby żaden z używanych w pracy zbiorów zmiennych

jednomodalnych i równomiernych, w szczególności dla zbioru {4,5} wartość miary (4.6) równa jest 0,101. Innym uzasadnieniem decyzji bezwzględnej może być to, że dla jednego z wymuszeń, dla pary zmiennych {1,2} $WKO^7(1,2) = 0{,}51 > 0{,}8 \cdot WKL^7(1,2)$. Jeszcze innym sposobem uzasadnienia decyzji bezwzględnej jest ocena wizualna pary {2,3}, dla tej bowiem pary nie możemy porównywać *WKO* z *WKL* , ponieważ *WKL*(2,3)>0,8. Na rysunku 5.3 widoczna jest bardzo wyraźna struktura dwóch skupień.



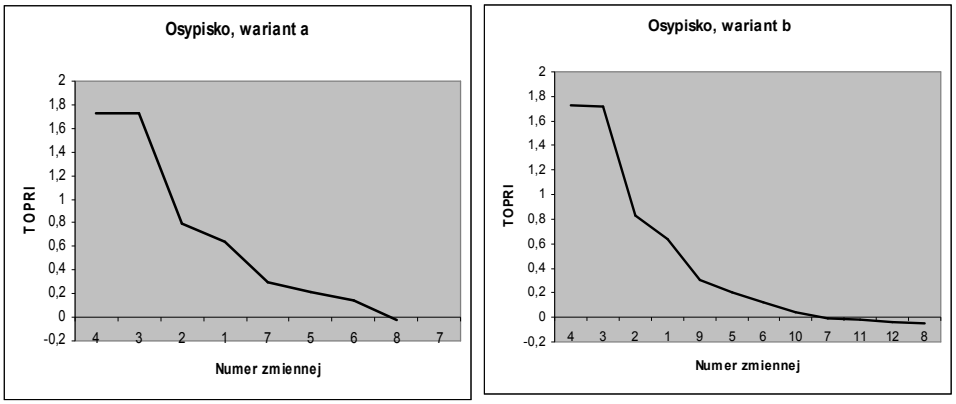
Rys. 5.3 Rzut zbioru *Invest* na 2 i 3 zmienną.

Źródło: Opracowanie własne.

Dla pary {1,2} ocena wizualna również daje jednoznaczny wniosek. Metoda *KO_ciągłe* w wariancie b) spisała się również bezbłędnie. Dla tego zbioru, o bardzo małej liczbie obiektów, wartości *WKL* oraz *WKO* były obliczane w wariancie kompletnym (por. uwaga 6 po def. 4.1).

Zbiór ***Iris***. Obiekty: kwiaty irysy. Liczebność: 150. Liczba zmiennych: 4. Liczba skupień: 3. Liczebności skupień: 50, 50, 50.
Charakterystyka skupień: 3 różne odmiany irysów.
Charakterystyka zmiennych: wszystkie zmienne są ciągłe i przedstawiają pomiary wybranych czterech elementów kwiatów.

Zbiór zaczerpnięty z repozytorium UCI, jest niemal nieodłącznym przykładem aplikacyjnym w większości prac z analizy skupień. Zbiór ten był przykładowo analizowany w kilku pracach dotyczących selekcji zmiennych w kontekście analizy skupień ale zawsze zbiór zmiennych maskujących strukturę skupień był bardzo prostej postaci. Zbadajmy efektywność wybranych metod selekcji zmiennych w dwóch wariantach.

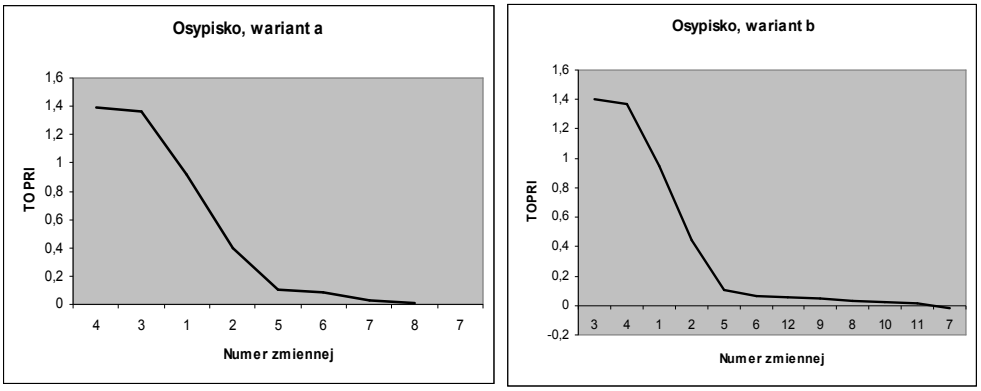


Rys. 5.4 Wykresy osypiska metody HINoV dla dwóch wariantów zbiorów zmiennych dla zbioru *Iris* dla liczby skupień równej 2.

Źródło: Opracowanie własne.

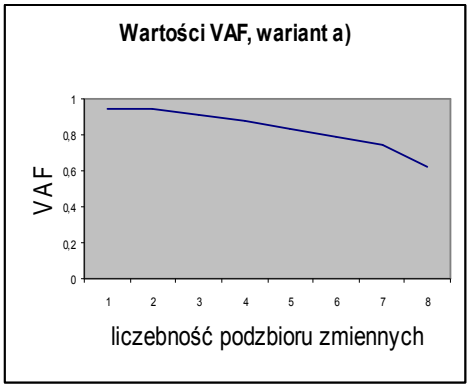
Wariant a) dwie zmienne (o numerach 5, 6) o dwuwymiarowym rozkładzie normalnym z zerowym wektorem średnim i macierzą kowariancji z jedynkami na głównej przekątnej i wartościami równymi 0,75 poza przekątną; trzecia zmienna (o numerze 7) o skośnym rozkładzie beta(1,1); czwarta zmienna (o numerze 8) o rozkładzie równomiernym na odcinku [0, 10]. Razem do analizy jest 8 zmiennych.

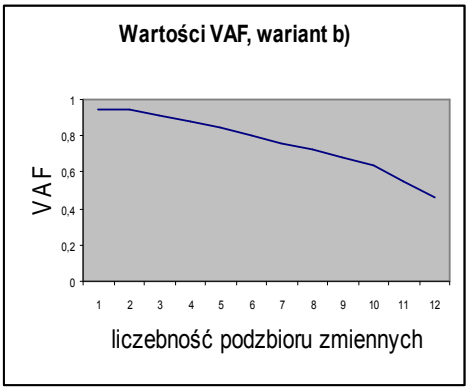
Wariant b) dwie zmienne (o numerach 5, 6) o dwuwymiarowym rozkładzie normalnym z zerowym wektorem średnim i macierzą kowariancji z jedynkami na głównej przekątnej i wartościami równymi 0,5 poza przekątną; dwie zmienne (o numerach 7, 8) o dwuwymiarowym rozkładzie normalnym z zerowym



Rys. 5.5 Wykresy osypiska metody HINoV dla dwóch wariantów zbiorów zmiennych dla zbioru *Iris* dla liczby skupień równej 3.

Źródło: Opracowanie własne.





Rys. 5.6 Wykresy wartości VAF dla metody VAF z indeksem skupialności, w dwóch wariantach zbiorów zmiennych, dla zbioru *Iris,* dla 3 skupień.

Źródło: Opracowanie własne.

wektorem średnim i macierzą kowariancji z jedynkami na głównej przekątnej i wartościami równymi 0,25 poza przekątną; dwie zmienne nieskorelowane (o numerach 9, 10) o skośnych rozkładach beta(1,1); dwie zmienne (o numerach 11, 12) o rozkładach równomiernych na odcinku [0, 10]. Razem do analizy jest 12 zmiennych.

Analiza replikacji zwracała niejednoznaczne wskazania w obu wariantach, częściej wskazywała, że najbardziej stabilnymi podziałami są podziały zbioru na 2 skupienia, rzadziej na 3. Wynik ten należy ocenić jako słaby – replikacja w przypadku tego zbioru danych nie odnajduje poprawnej liczby skupień choć liczba skupień nie ma dużego wpływu na wyniki selekcji zmiennych podane poniżej.

Metoda HINoV, z kryterium największego skoku, spisała się przeciętnie wybierając za każdym razem, niezależnie od liczby skupień, zmienne o numerach 3 i 4 – obie są zmiennymi istotnymi. Słuszność takiego wyboru potwierdza analiza wizualna wykresów osypiska i rysunki 5.4 i 5.5. Należy zauważyć, że we wszystkich wariantach metoda ustaliła poprawną kolejność wszystkich zmiennych w tym sensie, że żadna zmienna nieistotna nie ma wyższej rangi od żadnej zmiennej istotnej.

Metoda VSKM spisała się bardzo słabo w obu wariantach. W wariancie a) wybrała dwie zmienne o numerach 2 i 6 (dla liczby skupień równej 2) oraz zmienne o numerach 2 i 4 (dla liczby skupień równej 3). W wariancie b) również wybierała tylko po dwie zmienne: o numerach 1 i 7 (dla liczby skupień równej 2) oraz zmienne o numerach 5 i 8 (dla liczby skupień równej 3).

Metoda VAF z indeksem skupialności spisała tak jak HINoV. Analiza replikacji wskazała liczbę skupień równa 3. Dla tej liczby skupień metoda w obu

wariantach wybrała dwie zmienne o numerach 3 i 4 przy pomocy kryterium największego skoku. W obu wariantach metoda wybrała najsilniejsze dwie zmienne tworzące strukturę skupień. Te wybory, od strony wizualnej, są przedstawione na rysunku 5.6. Widoczny jest spadek wartości VAF po zbiorze o liczebności 2 zmiennych.

Metoda *KO_ciągłe* w wariancie a) spisała się bezbłędnie. W tabeli 5.2 podane są wartości współczynników korelacji liniowej i odległościowej dla

Tabela 5.2 Wartości *WKO, WKL* dla wybranych par zmiennych zbioru *Iris* w obu wariantach zmiennych zanieczyszczających.

Para zmiennych	*WKL*	*WKO*	WKL^7	WKO^7
{1, 2}	-0,13	0,00	-0,60	0,36
{1, 3}	0,87	0,66	0,92	0,87
{1, 4}	0,82	0,59	0,87	0,81
{2, 3}	-0,43	0,25	-0,83	0,62
{2, 4}	-0,35	0,23	-0,78	0,53
{3, 4}	0,96	0,91	0,98	0,98
{5, 6} war. a)	0,69	0,49	0,77	0,51
{5, 6} war. b)	0,52	0,18	0,57	0,22

Źródło: Opracowanie własne.

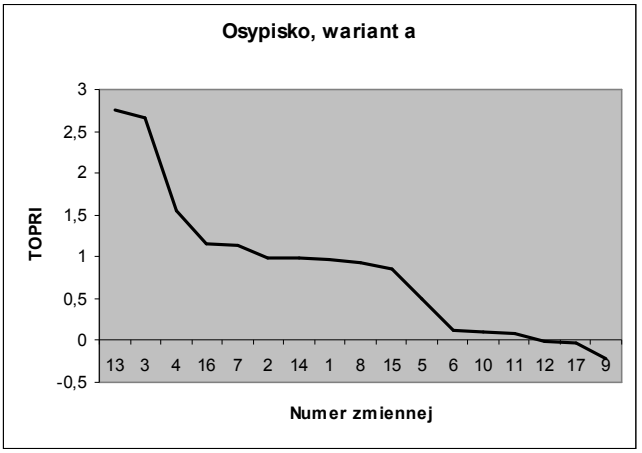
wybranych par zmiennych. Dla pozostałych par wszystkie wartości *WKL* oraz *WKO* były mniejsze od 0,13 co do wartości bezwzględnej. W pierwszym etapie, na bazie wartości *WKL* oraz *WKO,* metoda *KO_ciągłe* podzieliła zbiór wszystkich zmiennych na 4 podzbiory: {1,2,3,4}, {5,6}, {7} oraz {8}. W drugim etapie łatwo uzyskać decyzję bezwzględną dla podzbioru zmiennych {1,2,3,4} na podstawie wartości miary (4.6), która równą jest 1 (znacznie wyższą od 0,4). Wartość tę można interpretować jako doskonałą separowalność. Dla zbioru {5,6} wartość miary (4.6) równa jest 0,07.

W wariancie b) metoda *KO_ciągłe* również była bezbłędna. Nie pojawiła się żadna para zmiennych, która wprowadziłaby istotne zmiany do podziału zbioru wszystkich zmiennych. Został on podzielony na następujące podzbiory: {1,2,3,4}, {5,6}, {7}, {8}, {9}, {10}, {11}, {12}, para zmiennych {7,8} słabo skorelowanych nie została połączona w jeden podzbiór. Dalsza selekcja zmiennych przebiega tak samo jak w wariancie i otrzymujemy bezbłędną decyzję bezwzględną. Dla tego zbioru, o bardzo małej liczbie obiektów, wartości

WKL oraz *WKO* były obliczane w wariancie kompletnym (por. uwaga 6 po def. 4.1).

Zbiór ***Glass*** . Obiekty: próbki szkła. Liczebność: 214. Liczba zmiennych: 9. Liczba skupień: 6. Liczebności skupień: 76, 70, 29, 17, 13, 9.
Charakterystyka skupień: 6 różnych rodzajów szkła, od budowlanego okiennego, przez budowlane samochodowe do szkła kuchennego.
Charakterystyka zmiennych: wszystkie zmienne są ciągłe, pierwsza zmienna to współczynnik załamania światła, który daje próbka, osiem pozostałych to zawartości ośmiu pierwiastków: sodu, magnezu, aluminium, krzemu, potasu, wapnia, baru, żelaza.

Zbiór zaczerpnięty z repozytorium UCI, jest ciekawym zbiorem, który jest przykładem zastosowania analizy skupień w technice kryminalistycznej. Rodzaj szkła właściwie zidentyfikowany może być materiałem dowodowym. Zbadajmy efektywność wybranych metod selekcji zmiennych najpierw w jednym wariancie.



Rys. 5.7 Wykres osypiska metody HINoV dla wariantu a) zbioru zmiennych dla zbioru *Glass* dla liczby skupień równej 2.

Źródło: Opracowanie własne.

Wariant a) dwie zmienne (o numerach 10, 11) o dwuwymiarowym rozkładzie normalnym z zerowym wektorem średnim i macierzą kowariancji z jedynkami na głównej przekątnej i wartościami równymi 0,5 poza przekątną; dwie zmienne (o numerach 12, 13) o dwuwymiarowym rozkładzie normalnym z zerowym wektorem średnim i macierzą kowariancji z jedynkami na głównej przekątnej i wartościami równymi 0,25 poza przekątną; dwie zmienne

nieskorelowane (o numerach 14, 15) o skośnych rozkładach beta(1,1); dwie zmienne (o numerach 16, 17) o rozkładach równomiernych na odcinku [0, 10].

Tabela 5.3 Wartości *WKO, WKL* oraz dla par zmiennych zbioru *Glass.*

Para zmiennych	{1, 2}	{1, 3}	{1, 4}	{1, 5}	{1, 6}	{1, 7}	{1, 8}	{1, 9}
WKL	-0,11	-0,12	-0,45	-0,53	-0,33	0,76	-0,06	0,11
WKO	0,21	0,15	0,3	0,43	0,22	0,59	0,05	0,01
WKL^7	-0,02	-0,12	-0,67	-0,48	-0,67	0,83	-0,31	0,10
WKO^7	0,42	0,24	0,43	0,50	0,53	0,72	0,26	0,14
Para zmiennych	{2, 3}	{2, 4}	{2, 5}	{2, 6}	{2, 7}	{2, 8}	{2, 9}	
WKL	-0,28	0,15	-0,16	-0,42	-0,17	0,43	-0,22	
WKO	0,33	0,25	0,23	0,28	0,18	0,34	0,04	
WKL^7	-0,88	0,24	-0,44	-0,54	-0,12	0,60	-0,06	
WKO^7	0,89	0,38	0,50	0,76	0,33	0,61	0,06	
Para zmiennych	{3, 4}	{3, 5}	{3, 6}	{3, 7}	{3, 8}	{3, 9}		
WKL	-0,49	-0,21	0,22	-0,45	-0,48	0,11		
WKO	0,35	0,17	0,32	0,31	0,29	0,00		
WKL^7	-0,41	0,14	0,99	-0,44	-0,50	0,57		
WKO^7	0,77	0,18	0,98	0,52	0,34	0,32		
Para zmiennych	{4, 5}	{4, 6}	{4, 7}	{4, 8}	{4, 9}			
WKL	0,05	0,21	-0,25	0,53	-0,08			
WKO	0,24	0,38	0,14	0,32	-0,02			
WKL^7	0,15	0,68	-0,46	0,46	-0,49			
WKO^7	0,34	0,54	0,34	0,27	0,20			
Para zmiennych	{5, 6}	{5, 7}	{5, 8}	{5, 9}	{6, 7}	{6, 8}	{6, 9}	
WKL	-0,05	-0,21	-0,01	-0,08	-0,40	-0,17	0,05	
WKO	0,24	0,27	0,12	0,00	0,23	0,27	-0,05	
WKL^7	-0,06	-0,36	0,18	0,16	-0,49	-0,17	-0,10	
WKO^7	0,45	0,40	0,43	0,11	0,51	0,63	0,02	
Para zmiennych	{7, 8}	{7, 9}	{8, 9}					
WKL	-0,14	0,14	-0,06					
WKO	0,09	0,02	-0,05					
WKL^7	0,10	0,26	0,31					
WKO^7	0,42	0,11	0,28					

Źródło: Obliczenia własne.

Razem do analizy jest 17 zmiennych. Wyniki selekcji zmiennych w wariancie a) są bardzo słabe. Selekcja przeprowadzona została dla liczby skupień równej 2

(dla grupowania metodą *k*-średnich), ale prawie równie często, analiza replikacji i stabilności wskazywały na inne liczby skupień (3 oraz 4).

Kryterium największego skoku w metodzie HINoV wybrało dwie zmienne o numerach 3 i 10. Wizualna analiza wykresu osypiska (rys. 5.7) potwierdza ten bardzo słaby wynik. Na domiar złego należy zauważyć, że zjawisko łokcia jest na tym wykresie podwójne.

Metoda VSKM spisała się bardzo słabo, wybrała dwie zmienne o numerach 5 i 10.

Metoda VAF z indeksem skupialności spisała się bardzo źle odrzucając w etapie wstępnym 8 (!) spośród 9 zmiennych istotnych dla struktury skupień w obu wariantach zbiorów zmiennych. Wskaźniki skupialności były bardzo słabe, wszystkie mniejsze od 0,27 , niektóre mniejsze od 0,2. Ten wynik należy ocenić jako bardzo słaby. Dalszych wyników tej metody nie przedstawiamy, gdyż nie ma to w takiej sytuacji żadnego sensu.

W świetle powyższych rezultatów nie ma sensu kontynuowanie badań dla innych wariantów zbiorów zmiennych.

Metoda *KO_ciągłe* spisała się bezbłędnie. W tabeli 5.3 podane są wartości współczynników korelacji liniowej i odległościowej dla wszystkich par zmiennych oryginalnych zbioru. W pierwszym etapie metoda *KO_ciągłe* podzieliła zbiór wszystkich zmiennych na następujące podzbiory: {1,2,3,4,5,6,7,8,9}, {10,11}, {12}, {13}, {14}, {15}, {16} oraz {17}. W drugim etapie łatwo otrzymujemy decyzję bezwzględną dla podzbioru {1,2,3,4,5,6,7,8,9}. Na przykład, wartość miary (4.6) równa jest 0,888 (znacznie wyższa od 0,4) co świadczy o bardzo dobrej separowalności skupień . Dla zbioru {10,11} wartość miary (4.6) równa jest 0,09. Innym sposobem pozwalającym na podjęcie decyzji bezwzględnej jest porównanie *WKL* z *WKL*, dla wielu par zmiennych, na przykład: {1,2}, {1,5}, {1,6}, {2,4}, {2,5}, {2,6}, {2,7}, {2,8}, {3,4}, {3,7}, {4,5}, {5,6}, {5,8}, {6,7}, {6,8}, {7,8}.

Zbiór ***Powietrze***. Obiekty: województwa polskie i cała Polska. Liczebność: 17. Liczba zmiennych: 6. Liczba skupień: 3. Liczebności skupień: 10, 6, 1.
Charakterystyka skupień: 3 homogeniczne grupy województw.
Charakterystyka zmiennych: wszystkie zmienne są ciągłe i przedstawiają pomiary gęstości dróg, zagęszczenia liczby samochodów oraz zanieczyszczenia powietrza czterema związkami chemicznymi.

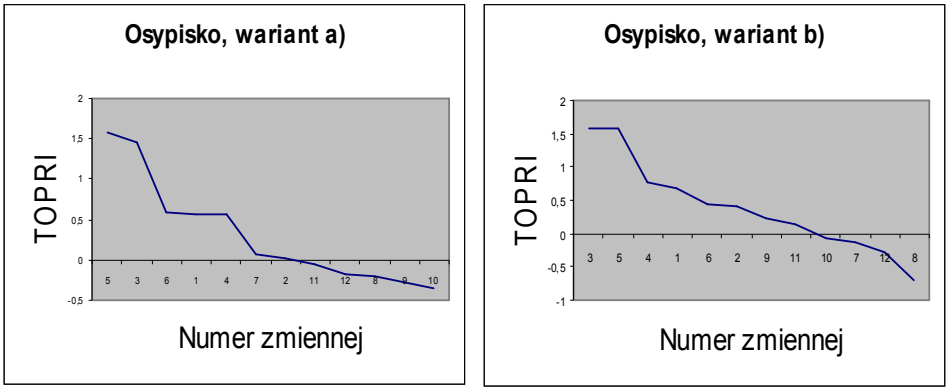
Zbiór zaczerpnięty z książki Walesiaka (2011). Zbadajmy efektywność wybranych metod selekcji zmiennych w dwóch wariantach.

Wariant a) dwie zmienne (o numerach 7, 8) o dwuwymiarowym rozkładzie normalnym z zerowym wektorem średnim i macierzą kowariancji z jedynkami na głównej przekątnej i wartościami równymi 0,75 poza przekątną;) dwie zmienne (o numerach 9, 10) o dwuwymiarowym rozkładzie normalnym z

zerowym wektorem średnim i macierzą kowariancji z jedynkami na głównej przekątnej i wartościami równymi 0,50 poza przekątną;) dwie zmienne (o numerach 11, 12) o rozkładzie równomiernym na odcinku [0, 10].

Wariant b) dwie zmienne (o numerach 7, 8) o dwuwymiarowym rozkładzie normalnym z zerowym wektorem średnim i macierzą kowariancji z jedynkami na głównej przekątnej i wartościami równymi 0,50 poza przekątną;) dwie zmienne (o numerach 9, 10) o dwuwymiarowym rozkładzie normalnym z zerowym wektorem średnim i macierzą kowariancji z jedynkami na głównej przekątnej i wartościami równymi 0,25 poza przekątną;) dwie zmienne (o numerach 11, 12) o rozkładzie równomiernym na odcinku [0, 10]. Razem, w każdym wariancie, do analizy jest 12 zmiennych.

Dla HINoV i jej modyfikacji można przeprowadzić replikację w ograniczonym zakresie tj. najwyżej do 8 skupień, ze względu na małą liczbę obiektów.



Rys. 5.8 Wykresy osypiska metody HINoV dla zbioru *Powietrze* dla liczby skupień równej 5.

Metoda HINoV spisała się słabo. W analizie replikacji najczęściej wskazywany był wariant 5 skupień. W wariancie a), przy metodzie największego skoku należy wybrać jako istotne tylko zmienne o numerach 3 i 5. Ilustracja tego wyboru jest widoczna po lewej stronie rysunku 5.8. Na wykresie osypiska widoczne jest dwukrotne zjawisko łokcia. Gdyby wybrać zmienne zgodnie z drugim łokciem, to otrzymalibyśmy o wiele lepszy wybór ale wyraźniejsze jest pierwsze zjawisko łokcia. Z kolei w wariancie b), zjawisko łokcia jest tylko jedno – należy wybrać dwie zmienne o numerach 3 i 5. Taki rezultat zwraca też metoda największego skoku.

Metoda VS-KM spisała się bardzo źle. W obu wariantach replikacja wskazywała prawie jednakowo często wszystkie rozważane liczby skupień. Zmienne wybierane były bardzo chaotycznie, na ogół 2 lub 3 ze zbioru

zmiennych istotnych, niekiedy dołączone były do nich 2 lub 3 zmienne ze zbioru zmiennych nieistotnych.

Tabela 5.4 Wartości *WKO, WKL* dla wybranych par zmiennych zbioru *Powietrze* w wariancie a) zmiennych zanieczyszczających.

Para zmiennych	{1,2}	{1,3}	{1,4}	{1,5}	{1,6}
WKL	0,07	0,78	0,74	0,72	0,69
WKO	-0,13	0,74	0,66	0,68	0,74
WKL^7	0,06	0,97	0,81	0,68	0,69
WKO^7	-0,07	0,75	0,71	0,66	0,73
Para zmiennych	{2,3}	{2,4}	{2,5}	{2,6}	
WKL	0,08	0,17	0,22	0,14	
WKO	-0,14	-0,02	-0,11	-0,14	
WKL^7	0,14	0,97	0,71	0,27	
WKO^7	-0,15	0,84	0,18	-0,14	
Para zmiennych	{3,4}	{3,5}	{3,6}		
WKL	0,88	0,91	0,89		
WKO	0,88	0,94	0,95		
WKL^7	0,91	0,96	0,96		
WKO^7	0,92	0,96	0,97		
Para zmiennych	{4,5}	{4,6}	{5,6}		
WKL	0,89	0,66	0,76		
WKO	0,87	0,80	0,88		
WKL^7	0,90	0,70	0,74		
WKO^7	0,87	0,86	0,90		
Para zmiennych	{7,8}	{9,10}			
WKL	0,74	0,22			
WKO	0,44	0,13			
WKL^7	0,79	0,55			
WKO^7	0,51	0,23			

Źródło: Obliczenia własne.

Metoda VAF z indeksem skupialności spisała się trochę lepiej niż HINoV . Jej wyniki są bardzo niestabilne, analiza wariancji wskazuje prawie wszystkie rozważane warianty liczby skupień. W niektórych przypadkach wynik jest dość dobry tzn. wybierane są (w obu wariantach a) oraz b)) zmienne {1,3,4,5,6}. Niekiedy jednak, podobnie jak dla HINoV, wybierane są tylko zmienne {3,5}.

Metoda *KO_ciągłe* spisała się bardzo dobrze. W tabeli 5.4 podane są wartości współczynników korelacji liniowej i odległościowej dla wszystkich par zmiennych oryginalnych zbioru oraz dla niektórych z pozostałych par z wariantu a) zmiennych nieistotnych. W wariancie a), w pierwszym etapie,, metoda *KO_ciągłe* podzieliła zbiór wszystkich zmiennych na następujące podzbiory: {1,2,3,4,5,6,}, {7,8}, {9,10}, {11}, {12}. W drugim etapie, decyzji bezwzględnej dla podzbioru {1,2,3,4,5,6} nie można oprzeć na mierze (4.6), gdyż dla tego zbioru nie można uzyskać grupowania metodą *k*-średnich na dwa skupienia – zawsze otrzymujemy tylko jedno skupienie. Można otrzymać decyzję bezwzględną na podstawie porównania *WKO* z *WKL* dla wielu par zmiennych, na przykład: {1,3}, {1,4}, {1,5}, {1,6}, {2,4}, {4,6}, {5,6}. Selekcja zmiennych przebiega analogicznie w wariancie b), w którym również otrzymujemy bezbłędną decyzję bezwzględną. Dla tego zbioru, o bardzo małej liczbie obiektów, wartości *WKL* oraz *WKO* były obliczane w wariancie kompletnym (por. uwaga 6 po def. 4.1).

Zbiór ***Australiancredit.*** Obiekty: posiadacze kart kredytowych. Liczebność: 690. Liczba zmiennych: 14. Liczba skupień: 2. Liczebności skupień: 383, 307.
Charakterystyka skupień: 2 grupy posiadaczy (gorsi i lepsi klienci).
Charakterystyka zmiennych: 8 zmiennych ciągłych, 6 zmiennych nominalnych o różnych liczbach wariantów (2, 2, 2, 3, 9, 14).

Zbiór zaczerpnięty z repozytorium UCI.

Dołączamy 14 zmiennych zakłócających o następujących rozkładach.

15, 16 – zmienne ciągłe silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,75 poza przekątną;

17, 18 – zmienne ciągłe dość silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,5 poza przekątną;

19, 20 – zmienne ciągłe słabo skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,25 poza przekątną;

21, 22 – zmienne ciągłe nieskorelowane, rozkłady beta, jeden stopień swobody dla licznika i jeden dla mianownika;

23, 24 – zmienne ciągłe nieskorelowane, rozkłady równomierne na odcinku [0;20];

25, 26 – zmienne porządkowe nieskorelowane, rozkłady równoliczne trzech wartości: 1, 2, 3;

27, 28 – zmienne binarne nieskorelowane, rozkłady równoliczne dwóch wartości.

Do analizy otrzymujemy zbiór 28 zmiennych różnych mierzonych na różnych skalach pomiarowych. Żadna z dotychczas opracowanych metod nie nadaje się do analizy tego zbioru danych. Wyniki otrzymane przez metodę *KO_meta* przedstawione są w tabelach 5.5-5.7.

Tabela 5.5 Wyniki metody *KO_meta* dla zbioru *Australiancredit* w wariancie 14 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{15,16}	**0,496**	**{9,10}**	**0,408**	**{2,7}**	**0,238**	**{7,10}**	**0,234**	**{3,7}**	**0,184**
{13,26}	0,039	{8,13}	0,203	{3,10}	0,262	{2,9}	0,300	{2,10}	0,306
{12,18}	0,053	{7,14}	0,128	{8,9}	0,171	{3,13}	0,202	{6,9}	0,160
{17,22}	0,069	{2,3}	0,156	{6,22}	0,115	{6,8}	0,154	{5,8}	0,165
{21,23}	0,073	{6,24}	0,133	{13,14}	0,102	{5,18}	0,085	{13,14}	0,071
{4,25}	0,054	{5,20}	0,064	{5,24}	0,069	{14,19}	0,057	{18,20}	0,049
{8,14}	0,046	{19,22}	0,068	{1,19}	0,04	{17,20}	0,066	{17,19}	0,055
{7,9}	0,082	{1,28}	0,044	{18,27}	0,035	{15,21}	0,048	{1,11}	0,039
{3,10}	0,113	{11,17}	0,035	{20,21}	0,045	{16,28}	0,054	{21,25}	0,024
{2,6}	0,053	{15,18}	0,039	{11,17}	0,041	{1,26}	0,028	{4,23}	0,035
{5,20}	0,062	{25,26}	0,046	{15,26}	0,029	{23,24}	0,022	{22,26}	0,025
{1,11}	0,026	{16,23}	0,041	{16,28}	0,042	{11,27}	0,026	{15,27}	0,015
{19,24}	0,015	{4,21}	0,017	{12,23}	0,041	{12,22}	-0,01	{24,28}	0,028
{27,28}	0,004	{12,27}	0,00	{4,25}	-0,02	{4,25}	-0,01	{12,16}	-0,02

Źródło: Opracowanie własne.

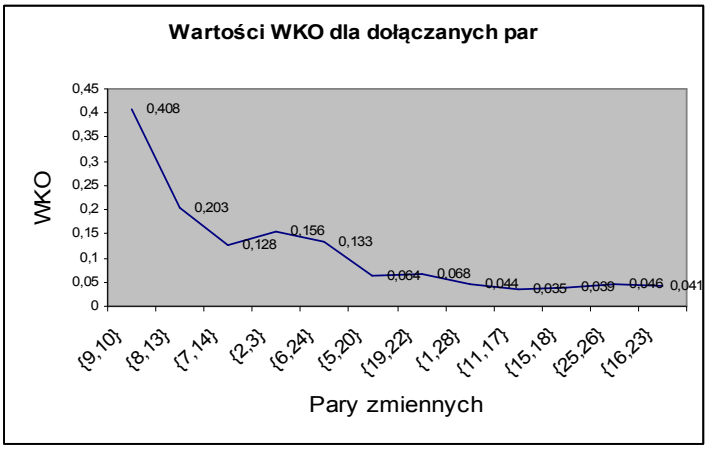
W tabeli 5.5 (i w następnych tabelach tego paragrafu), ciemniejszym kolorem zaznaczono miejsce zjawiska łokcia, tzn. krok, w którym należy przerwać dołączanie następnych par zmiennych, bo wartość *WKO* jest zbyt niska (mniejsza od 0,1). Wartość *WKO* jest obliczana pomiędzy parą bieżącą zmiennych, podaną w lewej kolumnie, a zbiorem wszystkich zmiennych dotychczas połączonych tj. wszystkich zmiennych należących do par podanych w lewej kolumnie od początku kolumny do wiersza poprzedzającego miejsce bieżące włącznie. Gdy wartość *WKO* jest zbyt niska, to nie ma sensu dołączanie nowych par do zbioru składającego się z par wcześniej połączonych, bo pomiędzy parą bieżącą a zbiorem par już połączonych nie ma związku w sensie korelacji odległościowej. Nie ulega wątpliwości, że metoda *KO_meta* spośród 28 zmiennych wybrała zmienne o numerach {2,3,5,6,7,8,9,10,14}. Te selekcje zostały przykładowo pokazane na rysunkach 5.9 oraz 5.10. Szczególnie sugestywny jest rysunek 5.10, który przedstawia wykres korelacji odległościowych wtedy, gdy łączenie zmiennych rozpoczynamy od pary {10,14}. Na początku tego ciągu par widoczny jest bardzo duży wzrost wartości

WKO pomiędzy dwiema parami zmiennych w porównaniu z *WKO* dla pary początkowej. Taka sytuacja zdarza się rzadko, ponieważ, na ogół (por. § 4.1),

Tabela 5.6 Wyniki metody *KO_meta* dla zbioru *Australiancredit* w wariancie 14 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{8,9}	**0,171**	**{17,18}**	**0,17**	**{10,14}**	**0,169**	**{2,3}**	**0,152**	**{3,10}**	**0,142**
{10,14}	0,271	{3,21}	0,067	{3,9}	0,410	{7,10}	0,283	{7,14}	0,239
{3,5}	0,095	{7,10}	0,129	{2,7}	0,209	{6,9}	0,161	{2,9}	0,248
{7,19}	0,124	{2,13}	0,156	{6,8}	0,172	{8,14}	0,168	{6,18}	0,147
{2,13}	0,162	{6,9}	0,155	{5,18}	0,095	{5,13}	0,090	{8,21}	0,13
{6,17}	0,144	{5,8}	0,112	{13,19}	0,054	{18,28}	0,044	{5,17}	0,107
{18,20}	0,079	{14,19}	0,069	{1,20}	0,059	{17,19}	0,057	{1,13}	0,043
{1,28}	0,057	{20,24}	0,051	{17,22}	0,026	{20,22}	0,045	{19,27}	0,036
{21,24}	0,039	{11,15}	0,029	{15,26}	0,029	{11,23}	0,045	{25,26}	0,025
{26,27}	0,025	{16,22}	0,049	{16,21}	0,051	{16,26}	0,022	{20,23}	0,056
{15,25}	0,024	{23,26}	0,038	{23,24}	0,038	{15,21}	0,088	{11,15}	0,027
{16,23}	0,024	{25,28}	0,033	{25,27}	0,026	{1,4}	0,032	{16,28}	0,026
{12,22]	0,001	{12,27}	0,025	{11,28}	0,022	{24,27}	0,008	{4,24}	0,011
{4,11}	-0,00	{1,4}	-0,01	{4,12}	-0,01	{12,25}	-0,01	{12,22}	-0,04

Źródło: Opracowanie własne.

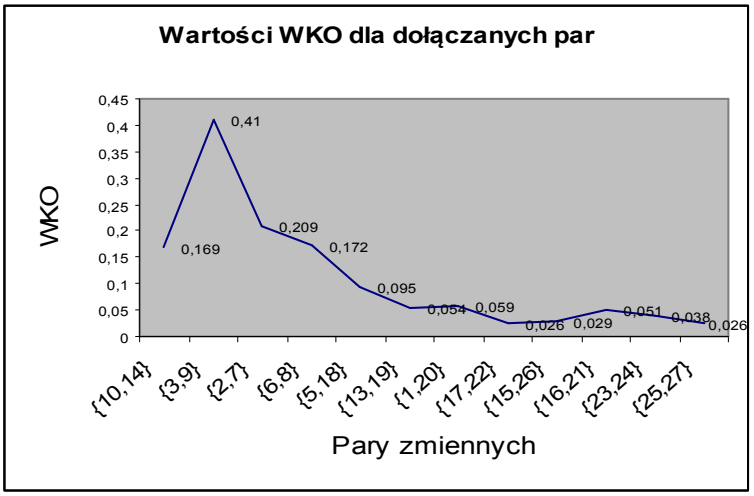


Rys. 5.9 Wykres wartości *WKO* pomiędzy parą bieżącą w ciągu par a zbiorem początkowych zmiennych od początku ciągu par, dla pary początkowej {9,10}, zbiór *Australiancredit*.
Źródło: Opracowanie własne.

Tablela 5.7 Wyniki metody *KO_meta* dla zbioru *Australiancredit* w wariancie 14 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{6,7}	**0,121**	**{2,10}**	**0,115**	**{2,6}**	**0,111**	**{5,6}**	**0,109**	**{3,14}**	**0,106**
{2,5}	0,189	{3,7}	0,315	{7,13}	0,174	{2,7}	0,13	{7,10}	0,223
{3,13}	0,15	{6,9}	0,203	{3,10}	0,258	{3,18}	0,148	{2,13}	0,253
{10,14}	0,211	{5,8}	0,167	{9,14}	0,141	{10,13}	0,174	{6,9}	0,167
{9,19}	0,129	{13,14}	0,073	{8,20}	0,142	{9,19}	0,131	{8,18}	0,15
{8,18}	0,122	{18,20}	0,057	{5,18}	0,076	{8,20}	0,133	{5,17}	0,077
{1,26}	0,045	{17,19}	0,073	{17,19}	0,075	{14,21}	0,072	{1,25}	0,055
{20,27}	0,037	{1,27}	0,059	{1,24}	0,04	{1,28}	0,039	{11,19}	0,034
{15,17}	0,034	{11,28}	0,023	{15,28}	0,035	{11,27}	0,031	{20,23}	0,034
{16,21}	0,051	{25,26}	0,025	{16,21}	0,044	{17,24}	0,042	{24,26}	0,041
{23,24}	0,032	{12,23}	0,016	{11,23}	0,023	{22,26}	0,029	{27,28}	0,018
{4,22}	0,013	{21,22}	0,017	{22,25}	0,011	{4,16}	0,03	{4,16}	0,013
{11,12}	0,031	{16,24}	0,012	{12,26}	0,00	{15,23}	0,045	{15,22}	0,06
{25,28}	-0,00	{4,15}	0,046	{4,27}	-0,02	{12,25}	-0,02	{12,21}	0,018

Źródło: obliczenia własne.



Rys. 5.10 Wykres wartości *WKO* pomiędzy parą bieżącą w ciągu par a zbiorem początkowych zmiennych od początku ciągu par , dla pary początkowej {10,14}, zbiór *Australiancredit.*
Źródło: Opracowanie własne.

wraz ze wzrostem liczebności zbiorów zmiennych maleje zdolność dyskryminacyjna *WKO*. Takie zjawisko można interpretować jako silne wskazanie związku pomiędzy obiema parami zmiennych, którego źródłem musi być tworzenie tej samej struktury skupień. Spośród wszystkich zmiennych oryginalnych pominięte zostały zmienne o numerach 1, 4, 11, 12. Zmienną o numerze 13 należy dołączyć do zbioru zmiennych wybranych ze względu na to, że kilkakrotnie znalazła się w zbiorze zmiennych znajdujących się przed łokciem, poprzedzając inne zmienne zdecydowanie istotne dla struktury skupień, na przykład, w ciągu par rozpoczętym przez parę {6,7} lub parę {2,6} (por. tab. 5.7). Zmiennych o numerach 17, 18, 19, 22, 24, które kilkakrotnie zostały dołączone nieznacznie przed progiem 0,1 nie włączamy do zbioru zmiennych wybranych ponieważ nie można potwierdzić dla żadnej z tych zmiennych oddzielnie, wysokiego skorelowania z żadnym podzbiorem innych zmiennych wybranych. Wynik w postaci zbioru {2,3,5,6,7,8,9,10,13,14} można uznać za dobry. Wyselekcjonowany zbiór zmiennych uznajemy w sposób bezwzględny za zbiór tworzący strukturę skupień, ze względu na to, że, na przykład, $WKO(\{2,3,7,10\},\{5,6,8,9\})=0{,}131>0{,}1$ i zmienne {5,6,8,9} to zmienne nominalne kilkustanowe. Ponadto, para zmiennych ciągłych {7,10} spełnia warunek pozwalający podjąć decyzję bezwzględną, gdyż $WKO(7{,}10)=0{,}242>0{,}8\cdot 0{,}300=0{,}8\cdot WKL(7{,}10)$.

Zbiór ***Votes.*** Obiekty: amerykańscy kongresmeni. Liczebność: 435. Liczba zmiennych: 16. Liczba skupień: 2. Liczebności skupień: 267, 168.
Charakterystyka skupień: 2 grupy kongresmenów (267 demokratów, 168 republikanów).

Charakterystyka zmiennych: 16 zmiennych binarnych reprezentujących wyniki głosowania (tak lub nie) nad 16 różnymi ustawami. Nieobecność kongresmena potraktowana została jako trzeci wariant ale nieobecności jest niewiele – dla wszystkich 16 zmiennych kilkanaście. Zbiór zaczerpnięty z repozytorium UCI.

Rozważmy dwa warianty zbiorów zmiennych nieistotnych. W wariancie a) dołączamy 14 zmiennych zakłócających.

17, 18 – zmienne ciągłe silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,75 poza przekątną;

19, 20 – zmienne ciągłe dość silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,5 poza przekątną;

21, 22 – zmienne ciągłe słabo skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,25 poza przekątną;

Tablica 5.8 Wyniki metody *KO_meta* dla zbioru *Votes* w wariancie 14 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{5,8}	**0,637**	**{5,9}**	**0,57**	**{4,5}**	**0,54**	**{7,8}**	**0,519**	**{3,4}**	**0,52**
{4,9}	0,688	{8,14}	0,658	{3,8}	0,68	{3,5}	0,664	{8,12}	0,64
{3,7}	0,655	{4,7}	0,639	{9,12}	0,649	{4,9}	0,717	{5,7}	0,64
{12,14}	0,576	{3,13}	0,604	{7,14}	0,603	{12,14}	0,567	{9,13}	0,63
{13,15}	0,519	{6,12}	0,547	{6,13}	0,497	{13,15}	0,515	{14,15}	0,49
{6,18}	0,365	{1,15}	0,394	{1,15}	0,390	{6,19}	0,372	{1,6}	0,37
{1,20}	0,212	{16,25}	0,145	{16,19}	0,118	{1,16}	0,233	{16,27}	0,12
{16,26}	0,119	{11,22}	0,091	{2,11}	0,069	{11,18}	0,088	{11,24}	0,08
{2,11}	0,079	{2,21}	0,073	{10,26}	0,055	{2,25}	0,061	{2,17}	0,07
{17,25}	0,053	{10,27}	0,066	{17,29}	0,043	{17,24}	0,051	{10,18}	0,07
{10,27}	0,049	{17,26}	0,029	{18,24}	0,059	{10,26}	0,052	{21,25}	0,03
{21,24}	0,031	{18,24}	0,051	{20,25}	0,035	{22,23}	0,039	{19,28}	0,03
{19,22}	0,032	{29,30}	0,016	{21,27}	0,030	{21,29}	0,029	{20,23}	0,04
{23,29}	-0,007	{19,23}	0,009	{28,30}	0,021	{27,30}	0,011	{22,26}	0,02
{28,30}	-0,003	{20,28}	-0,03	{22,23}	0,001	{20,28}	0,011	{29,30}	-0,01

Źródło: obliczenia własne.

23, 24 – zmienne ciągłe nieskorelowane, rozkłady beta, jeden stopień swobody dla licznika i jeden dla mianownika;

25, 26 – zmienne ciągłe nieskorelowane, rozkłady równomierne na odcinku [0;20];

27, 28, 29, 30 – zmienne binarne nieskorelowane, rozkłady równoliczne dwóch wartości.

Do analizy otrzymujemy zbiór 30 zmiennych różnych mierzonych na różnych skalach pomiarowych. Żadna z dotychczas opracowanych metod nie nadaje się do analizy tego zbioru danych. Wyniki otrzymane przez metodę *KO_meta.* Przedstawione są w tabelach 5.8-5.10.

Metoda *KO_meta* wybiera spośród 30 zmiennych wszystkie zmienne istotne tj. zmienne o numerach 1-16 oprócz zmiennych 2, 10. Ten wynik można uznać za dobry. Jednak , po dodatkowej analizie stwierdzamy, że obie brakujące zmienne również należy dołączyć do zbioru wybranych zmiennych. Na przykład, wymuszenie pary {2,6} jako pary startowej daje zawsze wynik dołączania par zmiennych ze zbioru {1, …,16}, na ogół pary {5,13} lub {5,14} przy wysokim

Tabela 5.9 Wyniki metody *KO_meta* dla zbioru *Votes* w wariancie 14 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{17,18}	**0,494**	**{4,8}**	**0,49**	**{3,8}**	**0,489**	**{8,9}**	**0,488**	**{5,7}**	**0,47**
{25,26}	0,055	{3,5}	0,771	{4,5}	0,675	{5,7}	0,699	{8,9}	0,70
{12,24}	0,059	{9,12}	0,648	{9,12}	0,653	{4,6}	0,602	{4,6}	0,58
{4,8}	0,141	{7,14}	0,633	{7,14}	0,622	{3,14}	0,625	{3,14}	0,61
{3,5}	0,635	{6,13}	0,482	{6,13}	0,488	{12,13}	0,575	{12,13}	0,58
{9,14}	0,61	{1,15}	0,394	{1,15}	0,398	{1,15}	0,393	{1,15}	0,41
{7,13}	0,58	{11,16}	0,13	{16,23}	0,13	{16,25}	0,125	{2,16}	0,14
{6,15}	0,356	{2,26}	0,068	{11,17}	0,092	{11,18}	0,08	{11,23}	0,07
{1,16}	0,214	{10,17}	0,06	{2,20}	0,073	{2,22}	0,068	{10,19}	0,04
{2,11}	0,1	{18,25}	0,037	{10,26}	0,079	{17,21}	0,065	{27,29}	0,03
{10,21}	0,035	{23,29}	0,025	{18,24}	0,048	{10,24}	0,06	{26,30}	0,02
{22,27}	0,045	{19,24}	0,02	{21,25}	0,042	{19,29}	0,032	{21,25}	0,03
{23,28}	0,014	{20,22}	0,047	{22,30}	0,029	{27,28}	0,011	{17,24}	0,02
{20,29}	0,013	{27,28}	0,02	{19,29}	0,009	{20,26}	0,035	{18,22}	0,05
{19,30}	0,026	{21,30}	-0,01	{27,28}	-0,01	{23,30}	-0,01	{20,28}	-0,01

Źródło: obliczenia własne.

WKO({2,6},{5,13})=0,31. Podobnie, do pary {4,11} jako startowej zawsze dołączane są pary zmiennych ze zbioru {1, …,16}, na ogół para {3,5} przy wysokim *WKO*({4,11},{3,5})=0,44. Podsumowując, stwierdzamy, że metoda *KO_meta* bezbłędnie, ze zbioru 30 zmiennych, wybiera zmienne o numerach {1, …,16}. Wyselekcjonowany zbiór uznajemy decyzją bezwzględną za zbiór tworzący strukturę skupień, ponieważ zawiera on zmienne dyskretne dwustanowe.

W wariancie b) dołączamy 8 zmiennych zakłócających, wszystkie są zmiennymi binarnymi o rozkładach równomiernych. Do analizy otrzymujemy zbiór 24 zmienne binarne. Wyniki otrzymane przez metodę *KO_meta* przedstawione są w tabelach 5.11-5.13. Godne odnotowania jest to, że przytaczane przykłady korelacji odległościowych dotyczą zmiennych binarnych i są bardzo silne w porównaniu z wartościami *WKO* jakie cechowały przykładową strukturę skupień badaną w punkcie 4.4.3 – dla tamtej struktury progiem była wartość 0,05. Decyzja jest taka sama jak w wariancie a), przy czym łatwiejsza do uzyskania – nie ma błędnych dołączeń.

Tabela 5.10 Wyniki metody *KO_meta* dla zbioru *Votes* w wariancie 14 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{4,12}	**0,448**	**{5,14}**	**0,433**	**{7,9}**	**0,422**	**{3,12}**	**0,42**	**{3,5}**	**0,398**
{3,5}	0,625	{4,8}	0,628	{5,8}	0,662	{4,8}	0,637	{4,8}	0,76
{8,14}	0,673	{3,9}	0,69	{3,4}	0,596	{5,7}	0,635	{7,9}	0,636
{7,9}	0,623	{7,12}	0,641	{12,13}	0,568	{9,13}	0,633	{12,13}	0,576
{6,13}	0,484	{13,15}	0,48	{14,15}	0,487	{14,15}	0,481	{14,15}	0,486
{1,15}	0,384	{1,6}	0,385	{6,22}	0,263	{6,22}	0,281	{6,25}	0,270
{16,19}	0,119	{16,17}	0,119	{1,16}	0,215	{1,23}	0,217	{1,21}	0,224
{11,26}	0,08	{11,18}	0,102	{11,27}	0,08	{16,17}	0,134	{16,17}	0,132
{2,10}	0,08	{2,21}	0,056	{2,10}	0,077	{2,10}	0,083	{11,18}	0,089
{23,29}	0,041	{10,27}	0,052	{23,26}	0,035	{11,24}	0,092	{2,22}	0,078
{17,22}	0,034	{25,26}	0,037	{17,24}	0,039	{21,26}	0,047	{10,20}	0,037
{18,28}	0,038	{24,29}	0,037	{18,25}	0,076	{25,27}	0,046	{19,24}	0,048
{24,25}	0,035	{19,20}	0,014	{21,29}	0,024	{18,19}	0,061	{28,29}	0,037
{20,30}	0,03	{23,28}	0,025	{20,28}	0,002	{28,30}	0,02	{26,30}	0,024
{21,27}	0,002	{22,30}	-0,02	{19,30}	-0,01	{20,29}	0,047	{23,27}	-0,001

Źródło: obliczenia własne.

Tabela 5.11 Wyniki metody *KO_meta* dla zbioru *Votes* w wariancie 8 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary Zm.	*WKO*	Pary Zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{5,8}	**0,622**	**{5,9}**	**0,56**	**{3,4}**	**0,548**	**{7,8}**	**0,53**	**{4,5}**	**0,52**
{4,9}	0,68	{4,8}	0,674	{8,12}	0,617	{5,9}	0,66	{3,8}	0,68
{3,7}	0,649	{3,7}	0,638	{5,7}	0,638	{3,4}	0,58	{9,12}	0,64
{12,13}	0,559	{12,13}	0,586	{9,13}	0,63	{12,14}	0,57	{7,14}	0,61
{14,15}	0,477	{6,14}	0,475	{14,15}	0,462	{13,15}	0,48	{13,15}	0,49
{1,6}	0,365	{1,15}	0,392	{1,6}	0,37	{1,6}	0,37	{1,6}	0,35
{16,17}	0,098	{11,16}	0,109	{2,16}	0,118	{10,16}	0,09	{11,16}	0,11
{2,23}	0,071	{2,19}	0,07	{10,11}	0,058	{11,19}	0,07	{2,10}	0,05
{11,18}	0,061	{10,21}	0,038	{17,19}	0,031	{2,18}	0,05	{22,24}	0,01
{10,21}	0,056	{18,24}	0,012	{21,23}	0,025	{21,24}	0,03	{17,21}	0,02
{20,24}	0,028	{17,20}	0,005	{20,22}	0,023	{17,20}	-0	{18,20}	0,01
{19,22}	-0,02	{22,23}	-0,01	{18,24}	-0,04	{22,23}	-0,01	{19,23}	0,00

Źródło: obliczenia własne.

Wszystkie zmienne o numerach wyższych od 16 (zmienne 17, 18, 19, 22, 23, 25, 26, 27) wykluczamy ze zbioru zmiennych wybranych (jeśli były wybrane), bo nie można potwierdzić oddzielnie dla żadnej z nich wysokiego skorelowania z żadnym podzbiorem zmiennych, które na pewno należy wybrać.

Tabela 5.12 Wyniki metody *KO_meta* dla zbioru *Votes* w wariancie 8 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{8,9}	**0,495**	**{4,8}**	**0,472**	**{3,8}**	**0,463**	**{5,7}**	**0,45**	**{5,14}**	**0,44**
{5,7}	0,718	{3,5}	0,744	{4,5}	0,673	{8,9}	0,71	{6,8}	0,62
{3,4}	0,586	{9,12}	0,649	{9,12}	0,669	{3,4}	0,59	{4,9}	0,65
{12,13}	0,571	{7,14}	0,615	{7,14}	0,603	{12,14}	0,58	{3,7}	0,62
{6,14}	0,465	{13,15}	0,508	{6,13}	0,506	{13,15}	0,48	{12,13}	0,59
{1,15}	0,378	{1,6}	0,355	{1,15}	0,405	{1,6}	0,38	{1,15}	0,41
{11,16}	0,108	{11,16}	0,118	{11,16}	0,118	{11,16}	0,12	{11,16}	0,12
{2,18}	0,053	{2,20}	0,049	{2,19}	0,062	{2,10}	0,07	{2,18}	0,06
{10,17}	0,048	{10,21}	0,032	{10,24}	0,038	{18,21}	0,02	{21,24}	0,06
{21,23}	0,053	{19,22}	0,02	{17,20}	0,044	{20,24}	0,03	{10,17}	0,04
{22,24}	0,032	{17,24}	0,004	{21,23}	0,034	{17,19}	0	{19,23}	0,02
{19,20}	0,015	{18,23}	0,01	{18,22}	-0,01	{22,23}	0,01	{20,22}	-0,01

Źródło: obliczenia własne.

Tabela 5.13 Wyniki metody *KO_meta* dla zbioru *Votes* w wariancie 8 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{4,12}	**0,418**	**{7,9}**	**0,413**	**{3,12}**	**0,402**	**(5,13)**	**0,39**	**(3,5)**	**0,375**
{3,5}	0,638	{5,8}	0,689	{4,8}	0,64	(4,8)	0,63	(4,8)	0,751
{8,9}	0,671	{4,6}	0,594	{5,7}	0,643	(3,9)	0,66	(9,12)	0,662
{7,14}	0,623	{3,14}	0,61	{9,14}	0,625	(7,12)	0,64	(7,14)	0,615
{6,13}	0,485	{12,13}	0,605	{6,13}	0,5	(14,15)	0,47	(13,15)	0,497
{1,15}	0,404	{1,15}	0,392	{1,15}	0,38	(6,16)	0,34	(1,6)	0,378
{2,16}	0,119	{11,16}	0,12	{11,16}	0,139	(1,11)	0,18	(11,16)	0,155
{11,21}	0,06	{2,24}	0,058	{2,10}	0,064	(2,21)	0,06	(2,10)	0,08
{10,19}	0,037	{10,21}	0,039	{23,24}	0,03	(23,24)	0,04	(21,23)	0,032
{22,23}	0,042	{19,23}	0,025	{17,22}	0,024	(10,22)	0,03	(18,19)	0,025
{17,24}	0,024	{18,20}	0,021	{19,21}	0,014	(17,19)	0,02	(17,20)	0,03
{18,20}	0,009	{17,22}	-0,03	{18,20}	-0,01	(18,20)	0,01	(22,24)	-0,01

Źródło: obliczenia własne.

Zbiór ***Wines.*** Obiekty: gatunki wina. Liczebność: 178. Liczba zmiennych: 13. Liczba skupień: 3. Liczebności skupień: 59, 71, 48.
Charakterystyka skupień: 3 gatunki wina.

Charakterystyka zmiennych: 13 zmiennych ciągłych reprezentujących zawartości 13 różnych składników win.

Zbiór zaczerpnięty z repozytorium UCI. Rozważmy następujący zbiór 12 zmiennych zakłócających.

14, 15 – zmienne ciągłe silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,75 poza przekątną;

16, 17 – zmienne ciągłe dość silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,5 poza przekątną;

18, 19 – zmienne ciągłe słabo skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,25 poza przekątną;

20, 21 – zmienne ciągłe nieskorelowane, rozkłady beta, jeden stopień swobody dla licznika i jeden dla mianownika;

22, 23 – zmienne ciągłe nieskorelowane, rozkłady równomierne na odcinku [0;20];

24, 25 – zmienne binarne nieskorelowane, rozkłady równoliczne dwóch wartości: 0, 1.

Tabela 5.14 Wyniki metody *KO_meta* dla zbioru *Wines* w wariancie 12 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{6,7}	**0,678**	**{7,12}**	**0,58**	**{14,15}**	**0,444**	**{6,12}**	**0,42**	**{7,9}**	**0,325**
{9,12}	0,589	{6,11}	0,618	{11,22}	0,08	{7,11}	0,658	{6,12}	0,634
{8,11}	0,352	{9,10}	0,383	{2,10}	0,165	{8,9}	0,377	{8,11}	0,371
{2,4}	0,3	{2,13}	0,291	{7,12}	0,234	{2,10}	0,302	{2,10}	0,314
{10,13}	0,209	{1,8}	0,276	{6,8}	0,376	{1,4}	0,145	{4,13}	0,154
{1,15}	0,169	{4,5}	0,159	{4,9}	0,243	{13,19}	0,193	{1,5}	0,159
{5,19}	0,099	{17,19}	0,052	{3,13}	0,134	{5,18}	0,114	{17,19}	0,055
{3,14}	0,081	{3,16}	0,039	{1,5}	0,146	{3,17}	0,044	{18,23}	0,047
{20,24}	0,039	{18,20}	0,038	{19,23}	0,046	{20,21}	0,038	{3,20}	0,034
{22,25}	0,029	{24,25}	0,036	{18,20}	0,025	{16,24}	0,044	{21,24}	0,037
{16,21}	0,041	{21,23}	0,027	{17,21}	0,015	{15,23}	0,022	{16,25}	0,012
{17,23}	0,022	{14,15}	0,003	{16,24}	0,005	{14,22}	0,045	{14,15}	0,007

Źródło: obliczenia własne.

Tabela 5.15 Wyniki metody *KO_meta* dla zbioru *Wines* w wariancie 12 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{11,12}	**0,322**	**{10,11}**	**0,312**	**{1,13}**	**0,301**	**{2,11}**	**0,275**	**{7,11}**	**0,269**
{7,10}	0,572	{7,12}	0,355	{4,6}	0,265	{10,12}	0,387	{6,12}	0,647
{2,6}	0,493	{6,8}	0,478	{5,7}	0,442	{6,7}	0,419	{8,9}	0,398
{8,9}	0,349	{2,9}	0,344	{9,12}	0,392	{8,9}	0,372	{2,10}	0,323
{4,13}	0,161	{4,13}	0,184	{8,11}	0,306	{1,4}	0,178	{1,4}	0,16
{1,17}	0,156	{1,19}	0,161	{2,10}	0,27	{13,17}	0,19	{3,13]}	0,174
{5,20}	0,086	{5,17}	0,081	{3,19}	0,055	{5,16}	0,07	{5,17}	0,074
{3,18}	0,042	{16,20}	0,048	{14,17}	0,044	{3,23]	0,043	{19,20}	0,068
{14,19}	0,036	{3,14}	0,036	{15,20}	0,067	{18,21}	0,032	{16,18}	0,046
{15,21}	0,053	{15,21}	0,071	{18,22}	0,03	{19,20}	0,05	{21,24}	0,01
{16,23}	0,06	{23,24}	0,001	{16,21}	0,029	{24,25}	0,02	{14,23}	0,044
{22,24}	0,018	{18,25}	0,005	{23,24}	0,023	{15,22}	-0,01	{15,22}	-0,00

Źródło: obliczenia własne.

Tabela 5.16 Wyniki metody *KO_meta* dla zbioru *Wines* w wariancie 12 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{6,9}	**0,267**	**{7,8}**	**0,262**	**{7,13}**	**0,261**	**{8,12}**	**0,248**	**{1,10}**	**0,232**
{5,7}	0,584	{6,12}	0,606	{1,6}	0,553	{7,11}	0,527	{11,13}	0,347
{12,13}	0,501	{9,11}	0,4	{9,12}	0,396	{6,10}	0,513	{4,7}	0,313
{1,8}	0,318	{2,10}	0,305	{8,11}	0,31	{2,9}	0,352	{6,12}	0,46
{4,11}	0,296	{1,4}	0,139	{4,10}	0,29	{4,13}	0,186	{8,9}	0,32
{2,10}	0,228	{3,13}	0,19	{2,5}	0,196	{1,5}	0,152	{2,5}	0,196
{16,19}	0,057	{5,17}	0,086	{3,17}	0,079	{17,19}	0,05	{3,19}	0,059
{15,17}	0,091	{18,23}	0,025	{16,20}	0,041	{3,16}	0,029	{17,20}	0,042
{14,20}	0,072	{19,20}	0,031	{18,19}	0,064	{18,23}	0,037	{15,25}	0,042
{23,24}	0,022	{14,21}	0,02	{15,24}	0,032	{20,24}	0,02	{14,18}	0,067
{3,18}	0,031	{16,25}	0,021	{14,22}	0,039	{22,25}	0,008	{16,21}	0,015
{21,25}	0,009	{15,22}	0,05	{21,23}	0,006	{14,21}	0,023	{22,24}	0,014

Źródło: obliczenia własne.

Do analizy otrzymujemy zbiór 25 zmiennych różnych mierzonych na różnych skalach pomiarowych. Żadna z dotychczas opracowanych metod nie nadaje się do analizy tego zbioru danych. Wyniki otrzymane przez metodę *KO_meta*. Przedstawione są w tabelach 5.14-5.16.

Metoda *KO_meta* wybiera spośród 25 zmiennych wszystkie zmienne istotne tj. zmienne o numerach {1, …,13} oprócz zmiennej 3 i 5. Ten wynik można uznać za dobry. Jednak, po dodatkowej analizie, zmienne 3 i 5 również należy dołączyć do zbioru wybranych zmiennych, gdyż, na przykład, *WKO*(3,4)=0,185>0,15 oraz *WKO*(3,5)=0,179>0,15. Ponadto, wymuszenie pary {3,5} jako pary startowej, daje zawsze wynik dołączania par zmiennych ze zbioru {1, …,13}, najpierw, na ogół, pary {4,9} przy *WKO*({3,5},{4,9})=0,180. Wszystkie zmienne o numerach wyższych od 13 (zmienne 15, 17, 18, 19) wykluczamy ze zbioru zmiennych wybranych (jeśli były wybrane), bo nie można potwierdzić oddzielnie dla żadnej z nich wysokiego skorelowania z żadnym podzbiorem zmiennych, które na pewno należy wybrać. Podsumowując, stwierdzamy, że metoda *KO_meta* bezbłędnie, ze zbioru 25 zmiennych, wyselekcjonowała zmienne o numerach {1, …,13}. Dla podzbioru {1,2,…,13} otrzymujemy decyzję bezwzględną na podstawie wartości miary (4.6) , która jest równa 0,406, czyli większa od 0,4.

Zbiór ***Ionosphere.*** Obiekty: dane z obserwacji teleskopowych jonosfery. Liczebność: 351. Liczba zmiennych: 34. Liczba skupień: 2. Liczebności skupień: 225, 126.
Charakterystyka skupień: „dobre” i „złe’ sygnały/dane.
Charakterystyka zmiennych: 34 zmienne ciągłe opisujące charakterystyki sygnałów. Zbiór zaczerpnięty z repozytorium UCI. Rozważmy następujący zbiór 16 zmiennych zakłócających.

35, 36 – zmienne ciągłe silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,75 poza przekątną;

37, 38 – zmienne ciągłe dość silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,5 poza przekątną;

39, 40 – zmienne ciągłe słabo skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,25 poza przekątną;

41, 42 – zmienne ciągłe nieskorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0 poza przekątną;

43, 44 – zmienne ciągłe nieskorelowane, rozkłady beta, jeden stopień swobody dla licznika i jeden dla mianownika;

45, 46 – zmienne ciągłe nieskorelowane, rozkłady równomierne na odcinku [0;20];

47, 48, 49, 50 – zmienne binarne nieskorelowane, rozkłady równoliczne dwóch wartości: 0, 1.

Do analizy otrzymujemy zbiór 50 zmiennych różnych mierzonych na różnych skalach pomiarowych. Żadna z dotychczas opracowanych metod nie nadaje się do analizy tego zbioru danych. Wyniki otrzymane przez metodę *KO_meta* przedstawione są w tabelach 5.17-5.19.

Tabela 5.17 Wyniki metody *KO_meta* dla zbioru *Ionosphere* w wariancie 16 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{13,15}	**0,727**	**{11,17}**	**0,613**	**{15,21}**	**0,567**	**{25,29}**	**0,556**	**{17,19}**	**0,542**
{17,21}	0,683	{13,19}	0,68	{13,19}	0,756	{27,31}	0,587	{11,21}	0,697
{11,19}	0,701	{9,15}	0,783	{7,17}	0,711	{23,33}	0,615	{15,23}	0,736
{9,23}	0,711	{10,21}	0,615	{9,23}	0,691	{18,21}	0,569	{9,13}	0,718
{7,25}	0,542	{7,23}	0,596	{11,25}	0,676	{14,19}	0,537	{7,25}	0,545
{10,29}	0,505	{8,25}	0,513	{10,29}	0,515	{16,17}	0,579	{10,33}	0,521
{8,33}	0,502	{12,29}	0,499	{12,33}	0,514	{12,15}	0,594	{8,29}	0,523
{12,31}	0,529	{6,33}	0,517	{8,31}	0,522	{10,11}	0,591	{6,31}	0,524
{6,27}	0,511	{4,31}	0,516	{6,27}	0,521	{13,20}	0,603	{5,12}	0,506
{5,14}	0,472	{5,14}	0,498	{4,5}	0,471	{6,9}	0,559	{4,27}	0,511
{4,18}	0,443	{22,27}	0,46	{14,22}	0,455	{7,8}	0,529	{14,20}	0,463
{16,22}	0,442	{20,28}	0,445	{3,16}	0,434	{5,22}	0,489	{16,22}	0,449
{20,28}	0,468	{16,26}	0,472	{20,32}	0,426	{4,24}	0,489	{18,28}	0,451
{24,26}	0,459	{3,18}	0,475	{18,28}	0,487	{26,28}	0,449	{24,26}	0,45
{1,30}	0,466	{1,24}	0,464	{24,30}	0,454	{1,30}	0,445	{30,32}	0,477
{3,32}	0,467	{30,32}	0,467	{26,34}	0,482	{3,34}	0,47	{1,34}	0,43
{2,34}	0,335	{2,34}	0,331	{1,2}	0,349	{2,32}	0,442	{2,3}	0,282
{36,42}	0,063	{36,37}	0,059	{36,41}	0,063	{41,42}	0,063	{36,41}	0,081
{45,49}	0,052	{41,45}	0,065	{42,48}	0,046	{36,40}	0,055	{42,45}	0,046
{37,41}	0,052	{47,50}	0,036	{45,50}	0,037	{45,50}	0,045	{46,49}	0,033
{46,50}	0,022	{48,49}	0,039	{40,47}	0,033	{37,46}	0,034	{37,38}	0,014
{47,48}	0,026	{42,44}	0,024	{35,49}	0,02	{38,47}	0,011	{35,47}	0,027
{35,40}	0,009	{35,46}	0,013	{44,46}	0,031	{44,49}	0,015	{39,48}	-0
{38,44}	-0,02	{38,40}	-0,01	{37,38}	-0,02	{43,48}	0,009	{40,50}	0,008
{39,43}	-0,05	{39,43}	-0,07	{39,43}	-0,07	{35,39}	0,025	{43,44}	-0,02

Źródło: obliczenia własne.

Ten zbiór ma tak silną strukturę skupień, że metoda *KO_meta* bezbłędnie wybiera dokładnie 34 zmienne tworzące strukturę skupień niezależnie od której pary rozpoczynamy łączenie zmiennych. Dla wyselekcjonowanego podzbioru

{1,2,…,34} łatwo otrzymać decyzję bezwzględną jeżeli zmodyfikujemy porównanie *WKO* z *WKL*. *WKO* pomiędzy drugą parą zmiennych a pierwszą parą jest w każdym ciągu dołączeń bardzo wysoki (większy od 0,5) natomiast *WKL* pomiędzy dowolną zmienną z drugiej pary i dowolną z pierwszej jest niski (mniejszy od 0,2). Wobec tego, tak wysoka wartość *WKO* musi być wynikiem istnienia struktury skupień. W oparciu o miarę (4.6) nie otrzymamy decyzji bezwzględnej, gdyż miara ta ma wartość 0,311, ale to wystarcza do tego, by pokonać w decyzji porównawczej dwa inne wyselekcjonowane podzbiory zmiennych {35,36} oraz {37,38}.

Tabela 5.18 Wyniki metody *KO_meta* dla zbioru *Ionosphere* w wariancie 16 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{15,17}	**0,542**	**{19,21}**	**0,541**	**{31,33}**	**0,535**	**{18,20}**	**0,531**	**{13,17}**	**0,524**
{13,19}	0,775	{15,17}	0,685	{7,29}	0,571	{22,30}	0,567	{11,15}	0,768
{11,21}	0,781	{13,23}	0,722	{5,25}	0,589	{16,26}	0,634	{9,21}	0,69
{9,23}	0,719	{11,33}	0,669	{3,21}	0,594	{4,24}	0,641	{7,19}	0,699
{7,25}	0,553	{9,25}	0,629	{9,23}	0,629	{14,28}	0,673	{10,23}	0,554
{10,33}	0,522	{7,29}	0,569	{15,19}	0,6	{12,32}	0,641	{8,25}	0,515
{8,31}	0,530	{10,31}	0,520	{11,13}	0,637	{6,34}	0,614	{12,33}	0,522
{6,29}	0,524	{6,27}	0,491	{17,27}	0,643	{8,10}	0,555	{6,29}	0,512
{5,12}	0,527	{5,12}	0,497	{6,10}	0,443	{1,23}	0,468	{4,31}	0,529
{4,27}	0,490	{8,14}	0,451	{8,12}	0,451	{21,31}	0,421	{14,27}	0,487
{14,22}	0,458	{4,22}	0,441	{4,14}	0,467	{19,33}	0,451	{5,20}	0,468
{3,16}	0,443	{3,20}	0,456	{1,20}	0,411	{15,25}	0,48	{16,22}	0,471
{20,28}	0,435	{16,26}	0,452	{16,22}	0,462	{13,29}	0,529	{3,18}	0,437
{18,26}	0,457	{18,24}	0,459	{18,24}	0,461	{7,17}	0,521	{28,30}	0,432
{24,30}	0,454	{28,30}	0,458	{26,28}	0,47	{9,27)}	0,495	{24,26}	0,482
{1,32}	0,460	{1,32}	0,470	{30,32}	0,464	{3,11}	0,463	{1,32}	0,452
{2,34}	0,359	{2,34}	0,341	{2,34}	0,357	{5,37}	0,299	{2,34}	0,343
{41,45}	0,058	{36,45}	0,06	{36,41}	0,069	{36,42}	0,064	{36,41}	0,084
{36,42}	0,038	{41,42}	0,065	{44,45}	0,042	{40,41}	0,057	{42,47}	0,043
{39,48}	0,025	{46,47}	0,037	{42,50}	0,030	{46,50}	0,037	{46,49}	0,026
{46,50}	0,031	{40,50}	0,025	{39,49}	0,030	{35,48}	0,03	{40,44}	0,03
{40,47}	0,036	{44,49}	0,017	{37,48}	0,037	{47,49}	0,029	{45,48}	0,045
{37,49}	0,008	{39,48}	0,003	{46,47}	0,000	{44,45}	0,038	{43,50}	0,003
{38,44}	-0,01	{37,43}	-0,01	{35,40}	0,002	{2,38}	0,011	{37,38}	-0,01
{35,43}	-0,02	{35,38}	-0,05	{38,43}	-0,03	{39,43}	0,074	{35,39}	-0,03

Źródło: obliczenia własne.

Tabela 5.19 Wyniki metody *KO_meta* dla zbioru *Ionosphere* w wariancie 16 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{9,11}	**0,522**	**{20,22}**	**0,514**	**{5,7}**	**0,514**	**{21,23}**	**0,504**	**{13,21}**	**0,502**
{15,17}	0,651	{18,24}	0,602	{3,13}	0,599	{17,33}	0,638	{15,17}	0,766
{13,19}	0,765	{16,26}	0,646	{9,15}	0,584	{15,31}	0,707	{11,19}	0,719
{21,23}	0,609	{14,28}	0,636	{11,21}	0,667	{13,29}	0,724	{9,23}	0,717
{7,10}	0,557	{6,30}	0,636	{17,33}	0,637	{11,25}	0,684	{7,25}	0,553
{6,25}	0,506	{12,32}	0,672	{19,23}	0,645	{7,19}	0,688	{10,33}	0,517
{8,31}	0,518	{4,34}	0,653	{10,25}	0,508	{9,27}	0,617	{8,31}	0,52
{12,33}	0,559	{1,8}	0,55	{8,29}	0,525	{5,10}	0,487	{6,29}	0,546
{4,29}	0,527	{10,31}	0,495	{12,31}	0,519	{3,8}	0,418	{5,12}	0,499
{14,27}	0,482	{21,33}	0,382	{6,27}	0,519	{4,12}	0,437	{4,27}	0,495
{5,22}	0,449	{13,29}	0,427	{4,14}	0,48	{6,14}	0,455	{3,14}	0,47
{16,28}	0,421	{15,23}	0,474	{16,22}	0,425	{22,28}	0,411	{16,28}	0,425
{20,26}	0,446	{7,19}	0,545	{1,20}	0,441	{16,20}	0,462	{22,30}	0,436
{18,24}	0,479	{11,25}	0,527	{18,28}	0,496	{18,30}	0,455	{20,26}	0,472
{30,32}	0,46	{5,17}	0,535	{24,26}	0,483	{24,32}	0,466	{18,24}	0,505
{3,34}	0,465	{9,27}	0,527	{30,32}	0,453	{26,34}	0,494	{1,32}	0,445
{1,2}	0,353	{2,3}	0,274	{2,34}	0,341	{1,2}	0,355	{2,34}	0,374
{41,42}	0,083	{36,41}	0,07	{36,42}	0,08	{36,41}	0,07	{36,41}	0,083
{36,45}	0,04	{35,42}	0,03	{45,50}	0,048	{47,50}	0,031	{40,45}	0,04
{47,50}	0,036	{45,49}	0,036	{41,44}	0,053	{45,46}	0,032	{35,42}	0,034
{37,48}	0,018	{44,46}	0,021	{46,49}	0,023	{48,49}	0,026	{38,50}	0,027
{35,46}	0,01	{40,48}	0,014	{37,47}	0,021	{40,42}	0,026	{39,49}	0,033
{39,49}	0,028	{37,47}	0,029	{43,48}	0,002	{38,44}	0,009	{37,48}	0,024
{38,40}	-0,01	{38,50}	-0,01	{35,38}	-0	{37,43}	0,012	{44,46}	0,007
{43,44}	-0,05	{39,43}	-0,05	{39,40}	-0,04	{35,39}	0,044	{43,47}	-0,02

Źródło: obliczenia własne.

Zbiór ***Concrete.*** Obiekty: próbki różnych rodzajów betonu. Liczebność: 1030. Liczba zmiennych: 8. Liczba skupień: 5. Liczebności skupień: zgodne z podziałem wytrzymałości betonu na 5 przedziałów o równych szerokościach.
Charakterystyka skupień: rodzaje betonu o danej wytrzymałości.
Charakterystyka zmiennych: 8 zmiennych ciągłych reprezentujących charakterystyki liczbowe betonu np. zawartość żużlu, wody itp..

Zbiór zaczerpnięty z repozytorium UCI.

Rozważmy następujący zbiór 8 zmiennych zakłócających.

9, 10 – zmienne ciągłe silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,75 poza przekątną;

11, 12 – zmienne ciągłe dość silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,5 poza przekątną;

13, 14 – zmienne ciągłe nieskorelowane, rozkłady równomierne na odcinku [0;20];

15, 16 – zmienne binarne nieskorelowane, rozkłady równoliczne dwóch wartości: 0, 1.

Do analizy otrzymujemy zbiór 16 zmiennych mierzonych na różnych skalach pomiarowych. Żadna z dotychczas opracowanych metod nie nadaje się do analizy tego zbioru danych. Wyniki otrzymane przez metodę *KO_meta* przedstawione są w tabeli 5.20. Łączenie zmiennych daje różne wyniki, najliczniejszy i jednocześnie bezbłędny podzbiór zmiennych {1,2,…,8} otrzymujemy zaczynając od pary {4,7}. Dołączenie ostatniej pary jest przy korelacji niewiele przekraczającej próg 0,1, ale nie było w żadnym przypadku błędnych dołączeń zmiennych zanieczyszczających. Nie można znaleźć potwierdzenia dla zmiennych 2 oraz 6 w postaci wysokiego skorelowania z innym podzbiorem zmiennych istotnych – w przypadku tego zbioru korelacje

Tabela 5.20 Wyniki metody *KO_meta* dla zbioru *Concrete* w wariancie 8 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{9,10}	**0,493**	**{4,5}**	**0,377**	**{4,7}**	**0,227**	**{3,5}**	**0,227**	**{11,12}**	**0,194**
{7,16}	0,032	{3,7}	0,203	{5,8}	0,273	{1,4}	0,232	{3,13}	0,037
{4,8}	0,086	{1,2}	0,189	{1,3}	0,126	{2,7}	0,108	{1,2}	0,153
{1,5}	0,130	{6,8}	0,079	{2,6}	0,108	{6,10}	0,088	{5,7}	0,141
{3,11}	0,128	{12,14}	0,041	{9,10}	0,093	{9,12}	0,089	{4,10}	0,149
{2,6}	0,087	{10,11}	0,021	{11,14}	0,053	{8,11}	0,033	{6,9}	0,084
{13,14}	0,029	{9,16}	0,029	{12,16}	0,003	{14,15}	0,006	{14,16}	0,020
{12,15}	-0,002	{13,15}	0,014	{13,15}	-0,028	{13,16}	-0,011	{8,15}	0,027

Źródło: obliczenia własne.

odległościowe nie są wysokie. Należy jednak zauważyć, że dzięki temu nie ma nieprawidłowych dołączeń zmiennych nieistotnych do zbioru zmiennych istotnych. Wyselekcjonowany podzbiór {1,2,…,8} ma bardzo wysoką wartość miary (4.6), równą 0,987, co od razu daje zdecydowaną decyzje bezwzględną o tym, że podzbiór ten tworzy strukturę skupień.

Zbiór ***Adult.*** Obiekty: dorosłe osoby wylosowane ze spisów. Liczebność: 1000 osób wylosowanych zależnie z 48842 osób. Liczba zmiennych: 14. Liczba skupień: 2. Liczebności skupień: zgodnie z wylosowaną próbą.
Charakterystyka skupień: osoby dobrze sobie radzące w życiu oraz osoby gorzej sobie radzące w życiu.
Charakterystyka zmiennych: 6 zmiennych ciągłych reprezentujących wiek, wagę, liczbę lat szkolnych, dochód finansowy, stratę finansową, liczbę godzin pracy w tygodniu oraz 8 zmiennych nominalnych reprezentujących kraj, stan cywilny, wykształcenie, płeć, rasę, zawód, rodzaj zatrudnienia, status rodzinny.
Zbiór zaczerpnięty z repozytorium UCI.

Rozważmy następujący zbiór 16 zmiennych zakłócających.
15, 16 – zmienne ciągłe silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,75 poza przekątną;
17, 18 – zmienne ciągłe dość silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,5 poza przekątną;
19, 20 – zmienne ciągłe słabo skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,25 poza przekątną;
21, 22 – zmienne ciągłe nieskorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0 poza przekątną;
23, 24 – zmienne ciągłe nieskorelowane, rozkłady beta, jeden stopień swobody dla licznika i jeden dla mianownika;
25, 26 – zmienne ciągłe nieskorelowane, rozkłady równomierne na odcinku [0;20];
27, 28, 29, 30 – zmienne binarne nieskorelowane, rozkłady równoliczne dwóch wartości: 0, 1.

Do analizy otrzymujemy zbiór 30 zmiennych mierzonych na różnych skalach pomiarowych. Żadna z dotychczas opracowanych metod nie nadaje się do analizy tego zbioru danych. Wyniki otrzymane przez metodę *KO_meta* przedstawione są w tabeli 5.21. W tabeli 5.21 przedstawione jest tylko łączenie zmiennych dla par startowych ze zbioru zmiennych oryginalnych. Łączenie daje różne wyniki, w zależności od pary startowej, ale po analizie łatwo widać bezbłędny rezultat, którym jest wybór wszystkich zmiennych ze zbioru {1,2,…,14}. Wątpliwości mogłaby budzić para {6,8} dołączona przy skorelowaniu minimalnie powyżej przyjętego progu, ale obawy te okazują się bezpodstawne, gdyż, na przykład, $WKO(6,10)=0,214$, albo $WKO(8,10)=0,363$ oraz $WKO(6,8)=0,552$.

Tabela 5.21 Wyniki metody *KO_meta* dla zbioru *Adult* w wariancie 16 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą parą zmiennych (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*	Pary zm.	*WKO*
{5,14}	**0,437**	**{4,5}**	**0,375**	**{5,13}**	**0,349**	**{13,14}**	**0,339**	**{2,13}**	**0,316**
{3,13}	0,409	{2,14}	0,311	{4,14}	0,548	{5,12}	0,449	{1,5}	0,320
{2,4}	0,411	{3,13}	0,460	{2,12}	0,296	{2,4}	0,390	{4,14}	0,522
{1,7}	0,188	{7,12}	0,190	{1,7}	0,194	{1,7}	0,196	{7,12}	0,214
{11,12}	0,129	{9,15}	0,119	{9,10}	0,132	{9,10}	0,127	{9,15}	0,114
{9,10}	0,104	{10,16}	0,111	{8,11}	0,073	{6,8}	0,101	{10,11}	0,103
{8,15}	0,078	{1,11}	0,094	{3,6}	0,132	{3,15}	0,033	{8,16}	0,067
{6,17}	0,087	{8,18}	0,062	{16,19}	0,017	{16,19}	0,024	{3,6}	0,106
{18,20}	0,031	{6,17}	0,087	{15,20}	0,033	{11,18}	0,024	{18,19}	0,019
{16,19}	-0,009	{19,20}	-0,003	{17,18	0,009	{17,20}	-0,030	{17,20}	-0,015

Źródło: obliczenia własne.

Zmiennych o numerach 15 i 16, również dołączonych nieznacznie powyżej progu 0,1, w parach z innymi zmiennymi, nie uwzględniamy gdyż nie ma skorelowania z żadnym innym podzbiorem zmiennych istotnych dla struktury skupień potwierdzającego przynależność do zbioru zmiennych {1,2,…,14}. Dla wyselekcjonowanego podzbioru od razu otrzymujemy decyzję bezwzględną, gdyż zbiór ten zawiera zmienne dyskretne kilkustanowe.

Zbiór ***Donation.*** Obiekty: dorosłe osoby będące dawcami krwi. Liczebność: 748. Liczba zmiennych: 4. Liczba skupień: 2. Liczebności skupień: 180, 568.
Charakterystyka skupień: osoby, które oddały w marcu 2007 oraz osoby, które nie oddały krwi w tym miesiącu.
Charakterystyka zmiennych: 4 zmienne ciągłe reprezentujące liczbę miesięcy od ostatniego oddania krwi, całkowitą liczbę oddań krwi, objętość do tej pory oddanej krwi, liczbę miesięcy od pierwszego oddania krwi.

Zbiór zaczerpnięty z repozytorium UCI. Rozważmy następujący zbiór 5 zmiennych zakłócających.

6, 7 – zmienne ciągłe silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,75 poza przekątną;

8, 9 – zmienne ciągłe nieskorelowane, rozkłady równomierne na odcinku [0;20];

10 – zmienna binarna, rozkład równomierny dwóch wartości: 0, 1.

Do analizy otrzymujemy zbiór 9 zmiennych mierzonych na różnych skalach pomiarowych. Żadna z dotychczas opracowanych metod nie nadaje się do

analizy tego zbioru danych. Wyniki otrzymane przez metodę *KO_meta* przedstawione są w tabeli 5.22. W tym przypadku, do pary początkowej dołączane były jednoelementowe zbiory zmiennych ze względu na małą liczbę wszystkich zmiennych. Z tabeli 5.22 wynika, że bardzo silne są wskazania połączenia zmiennych 2, 3 i 4. Do tych zmiennych dołączamy zmienną numer 1 ponieważ *WKO*({2,3},{1})=0,110>0,1. Wyselekcjonowane zostały zatem podzbiory zmiennych {1,2,3,4} oraz {5,6}. Dla podzbioru {1,2,3,4} otrzymujemy decyzje bezwzględną na podstawie bardzo wysokiej wartości miary (4.6), która równa jest 0,904.

Tabela 5.22 Wyniki metody *KO_meta* dla zbioru *Donation* w wariancie 5 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą zmienną (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Zmienne	*WKO*	Zmienne	*WKO*	Zmienne	*WKO*	Zmienne	*WKO*
{2,3}	**1,00**	**{5,6}**	**0,463**	**{2,4}**	**0,418**	**{3,4}**	**0,431**
{4}	0,411	{1}	0,038	{3}	0,533	{2}	0,552
{1}	0,020	{2}	0,088	{1}	0,025	{1}	0,027
{5}	0,010	{4}	0,074	{5}	0,030	{5}	0,027
{8}	-0,009	{7}	0,029	{8}	-0,018	{6}	0,002

Źródło: obliczenia własne.

Zbiór ***Nieruchomości.*** Obiekty: nieruchomości z województwa jeleniogórskiego. Liczebność: 27. Liczba zmiennych: 6. Liczba skupień: 4. Liczebności skupień: 9, 4, 5, 9.
Charakterystyka skupień: cztery homogeniczne grupy nieruchomości.
Charakterystyka zmiennych: 3 zmienne porządkowe reprezentujące lokalizację nieruchomości gruntowej, standard użytkowy lokalu, warunki bytowe nieruchomości gruntowej oraz 3 zmienne nominalne reprezentujące położenie nieruchomości gruntowej, typ wspólnoty mieszkaniowej, powierzchnię gruntu w stosunku do obrysu budynku.
Zbiór zaczerpnięty z książki Walesiaka (2011).

Rozważmy następujący zbiór 6 zmiennych zakłócających.

7, 8 – zmienne ciągłe silnie skorelowane, rozkład normalny, średnie równe 0, w macierzy kowariancji 1 na przekątnej oraz 0,75 poza przekątną;

9, 10 – zmienne ciągłe nieskorelowane, rozkłady równomierne na odcinku [0;20];

11, 12 – zmienne binarne nieskorelowane o rozkładach równomiernych dwóch wartości: 0, 1.

Do analizy otrzymujemy zbiór 12 zmiennych mierzonych na różnych skalach pomiarowych. Żadna z dotychczas opracowanych metod nie nadaje się do analizy tego zbioru danych. Wyniki otrzymane przez metodę *KO_meta* przedstawione są w tabeli 5.23. W tym przypadku, podobnie jak dla zbioru *Donation,* do pary początkowej dołączane były jednoelementowe zbiory zmiennych ze względu na małą liczbę wszystkich zmiennych. Ponadto, zastosowano kompletną metodę obliczania wartości *WKO* ze względu na małą liczebność zbioru danych. Z tabeli 5.23 wynika, że bardzo silne są wskazania połączenia zmiennych 1, 2, 3, 4 i 6. Do tych zmiennych należy jednak dołączyć zmienną numer 5 ponieważ , na przykład, *WKO*({2,3},{5})=0,125>0,1. Wybrany został zatem bezbłędny podzbiór zmiennych {1,2,3,4,5,6}. Podzbiór ten uznajemy za tworzący strukturę skupień decyzją bezwzględną, ponieważ zawiera on kilkustanowe zmienne nominalne i porządkowe.

Tabela 5.23 Wyniki metody *KO_meta* dla zbioru *Nieruchomości* w wariancie 6 zmiennych nieistotnych. Wartość *WKO* to współczynnik korelacji odległościowej pomiędzy bieżącą zmienną (w lewej kolumnie) a wszystkimi dotychczas połączonymi zmiennymi.

Zmienne	*WKO*	Zmienne	*WKO*	Zmienne	*WKO*	Zmienne	*WKO*
{2,3}	**0,393**	**{7,8}**	**0,358**	**{1,5}**	**0,304**	**{1,4}**	**0,261**
{6}	0,304	{11}	0,045	{3}	0,065	{6}	0,436
{4}	0,163	{6}	0,031	{2}	-0,123	{3}	0,243
{8}	0,006	{2,}	0,033	{7}	0,023	{2}	0,216
{9}	0,015	{3}	0,001	{9}	-0,044	{8}	0,035
{10}	-0,040	{1}	-0,030	{10}	0,017	{5}	-0,112
{1}	-0,021	{5}	0,021	{6}	-0,020	{10}	0,011
{5}	0,012	{4}	0,001	{4}	0,011	{7}	-0,018

Źródło: obliczenia własne.

5.4. Ocena efektywności aplikacyjnej metod selekcji zmiennych

Przeprowadzone badania o charakterze aplikacyjnym na kilkunastu empirycznych zbiorach danych pozwalają na sformułowanie następującej oceny stosowanych metod.

1. Wśród dotychczas skonstruowanych metod, które można znaleźć w literaturze przedmiotu, brakuje metody, która jest najbardziej potrzebna – metody, przy pomocy której można by analizować dane mierzone na skalach o różnej precyzji pomiaru. Metoda *KO_meta* jest jedyną efektywną metodą, którą można stosować w przypadku takich zbiorów.
2. Jeśli problem selekcji zmiennych ograniczymy do zbiorów ze danymi ciągłymi, to spośród dotychczas opracowanych metod, najlepsze w eksperymentach symulacyjnych metody tj HINoV, VS-KM oraz VAF z indeksem skupialności osiągnęły słabe wyniki. Wszystkie metody, a zwłaszcza modyfikacje HINoV okazują się być opracowanymi pod kątem eksperymentów, w których zbiory danych generowane są w postaci mieszanin rozkładów normalnych. W przypadku empirycznych zbiorów danych, na przykład, indeks skupialności, który miał pomagać w odrzucaniu zmiennych nieistotnych we wstępnej selekcji, odrzuca zbyt często zmienne istotne dla struktury skupień. HINoV oraz VS-KM są odporne na skorelowanie zmiennych zakłócających tylko wtedy, gdy struktura skupień jest bardzo silna.
3. Metoda HINoV, która spośród dotychczas opracowanych metod wypadła najlepiej w zastosowaniu do empirycznych zbiorów danych, jest dużo słabsza od zaproponowanych w pracy procedur *KO_ciągłe* oraz *KO_meta* . HINoV traci swą zdolność dyskryminacyjną, gdy liczba wszystkich zmiennych, z których należy dokonywać selekcji jest rzędu kilkunastu.
4. Metody *KO_ciągłe* oraz *KO_meta* spisały się dobrze lub bardzo dobrze w przypadku wszystkich zbadanych empirycznych zbiorów danych. Metoda *KO_meta* okazała się bardzo elastyczna. Jeżeli korelacja odległościowa pomiędzy różnymi podzbiorami zmiennych jest silna, to może wystąpić zjawisko dołączania zmiennych nieistotnych występujących w parach ze zmiennymi istotnymi. Ale w takich przypadkach można oddzielnie sprawdzić każdą ze zmiennych z pary pod kątem jej silnego skorelowania ze zmiennymi połączonymi wcześniej. Jeżeli korelacja odległościowa pomiędzy różnymi podzbiorami zmiennych nie jest silna, to nie powinno występować zjawisko błędnego dołączania zmiennych nieistotnych.
5. Metoda *KO_meta* działa bardzo efektywnie nawet w przypadku zbiorów z kilkudziesięcioma zmiennymi, z których około połowa to zmienne tworzące strukturę skupień.

Zakończenie

Badania przeprowadzone w rozprawie i uzyskane w nich wyniki pozwalają na wyciągnięcie następujących wniosków końcowych w odniesieniu do celów rozprawy sformułowanych we wstępie.

Ocena dotychczasowego dorobku nauki

1. Dotychczasowy dorobek nauki w temacie selekcji zmiennych w analizie skupień obejmuje kilkanaście opracowanych w tym celu metod. Metody te można podzielić na dwie grupy: metody modelowe, które niekiedy dają wyniki w postaci wnioskowania statystycznego oraz metody heurystyczne, które są algorytmami postępowania.
2. Metody modelowe nie zachowują swojej cechy podstawowej, ponieważ modelowanie zbiorów danych w postaci mieszaniny rozkładów normalnych dla zbiorów empirycznych nie zawsze jest uzasadnione. Metody te są wówczas tylko algorytmami postępowania.
3. Metody modelowe spisują się słabo w eksperymentach symulacyjnych, w których zbiory danych generowane są zgodnie z założeniem potrzebnym do funkcjonowania metody tj. w postaci mieszaniny rozkładów normalnych.
4. Najlepsze z dotychczas opracowanych metod dla zbiorów danych ciągłych, to HINoV i dwie modyfikacje tej metody: VSKM oraz metoda VAF z indeksem skupialności. Obie modyfikacje mają mniejszy zakres stosowalności od HINoV, gdyż nadają się tylko do zbiorów danych ze zmiennymi ciągłymi, natomiast HINoV można stosować do zbiorów danych ze zmiennymi, na przykład, porządkowymi. Ponadto, VS-KM spisuje się dobrze tylko w eksperymentach symulacyjnych, w zastosowaniu do empirycznych zbiorów danych okazał się słabszy od HINoV.
5. Najlepsze dotychczas opracowane metody wybierają zmienne jedynie w sposób porównawczy tzn. na zasadzie wybierania jednego (z dwóch) podzbioru zmiennych, który ma silniejsze cechy świadczące istnieniu struktury skupień niż zbiór pozostałych zmiennych.
6. Konsekwencją porównawczego charakteru dotychczas opracowanych metod jest to, że nigdy metody nie wykryją wielokrotnych struktur skupień.
7. Wszystkie dotychczas opracowane metody zachowują swoje własności tylko do liczby kilku zmiennych opisujących zbiory danych. Dla liczby zmiennych wyższej od 8 wszystkie te metody tracą swą efektywność selekcji zmiennych.
8. Najlepsze z dotychczas opracowanych metod są uzależnione od znajomości liczby skupień dla jakiej metoda zostanie zastosowana. To jest

istotną wadą tych metod, gdyż, w praktyce nie ma żadnych możliwości wyznaczenia liczby skupień w zbiorze danych przed wybraniem zmiennych opisujących obiekty. Nawet po wybraniu zmiennych to zadanie bardzo rzadko można precyzyjnie rozwiązać. Wadę uzależnienia od koniecznej znajomości liczby skupień można próbować obejść stosując badanie replikacyjne lub stabilnościowe, ale to wiąże się ze spadkiem efektywności metod.

9. Jedyną z dotychczas opracowanych metod, którą można zaadoptować do zbiorów danych ze zmiennymi mierzonymi na różnych skalach jest metoda Ichino. Metoda ta cechuje się jednak w tych przypadkach bardzo słabą efektywnością.

Ocena zaproponowanych nowych procedur

W rozprawie zostało zaproponowanych kilka nowych procedur opartych na współczynniku korelacji odległościowej. Ogólnie rzecz biorąc, podstawą selekcji zmiennych jest dodatnia korelacja odległościowa pomiędzy dwoma różnymi zbiorami zmiennych wnoszących wkład do struktury skupień zbioru obiektów. Ponadto, zaproponowane procedury wykorzystują wielostopniowe grupowanie obiektów w dwa skupienia zamiast grupowania w ustaloną liczbę skupień, którą należałoby znać. Metody te mają następujące cechy:

- do zastosowania metody nie potrzebna jest dokładna znajomość liczby skupień w zbiorze danych;
- metody, często podejmują decyzje w sposób bezwzględny a nie porównawczy;
- metody można niekiedy stosować do wybierania zmiennych tworzących wielokrotne struktury skupień;
- metody można stosować do zbiorów danych opisanych przez zmienne mierzone na różnych skalach pomiarowych.

Zaproponowana metoda *KO_ciągłe* spisała się bardzo dobrze w eksperymentach symulacyjnych, w których zbiory danych są generowane z mieszanin rozkładów normalnych. Efektywność jest wyższa od efektywności najlepszych dotychczas opracowanych metod selekcji zmiennych w analizie skupień.

Zaproponowana metoda *KO_ciągłe* spisała się bardzo dobrze również w zastosowaniu do empirycznych zbiorów danych pomimo to, że wartości progów zostały wyprowadzone dla przykładowych struktur skupień będących mieszankami rozkładów normalnych.

Zaproponowana metoda *KO_binarne* spisała się bardzo dobrze w eksperymencie symulacyjnym na przykładowej strukturze skupień. Efektywność tej metody jest wyższa od efektywności najlepszej spośród dotychczas opracowanych metod selekcji zmiennych tego typu.

Zaproponowana metoda *KO_meta* nie ma konkurencji w postaci efektywnej metody, którą można by zastosować w przypadku zbiorów danych, w których mamy do czynienia z występowaniem różnych skal pomiarowych zmiennych opisujących obiekty zbioru. Na zbadanych zbiorach danych empirycznych metoda spisała się bardzo dobrze.

Metody zostały skonstruowane na podstawie wskaźników otrzymanych z przykładowych struktur skupień wygenerowanych z mieszanek rozkładów normalnych oraz przykładowych zmiennych zakłócających w postaci wybranych rozkładów jednomodalnych i równomiernych, ale wydaje się, że są wystarczająco ogólne, by pomóc w prawidłowym przeprowadzeniu analizy skupień dla wielu zbiorów empirycznych.

Skonstruowane metody nie wyczerpują wszystkich możliwości adaptacyjnych zaproponowanych mierników i wskaźników. Można, na przykład, skonstruować metodę, która będzie wykorzystywała ideę wielostopniowego grupowania w dwa skupienia połączoną z inną miarą jakości otrzymanego grupowania. Tą miarą może być, na przykład, indeks Sokołowskiego zastosowany do zbioru odległości pomiędzy parami obiektów, i zmodyfikowany tak, by przyjmował wysoką wartość wtedy gdy rozkład ten będzie wyraźnie dwumodalny.

Można też próbować konstruować empiryczne testy istotności wykorzystujące współczynnik korelacji odległościowej.

Literatura

Agrawal R., Gehrke J., Gunopulos D., Raghavan P.(1998), *Automatic subspace clustering of high dimensional data for data mining applications* , Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data, Seattle, Washington.

Arabie P., Hubert L., (1994), *Cluster analysis in marketing research,* Advanced methods of marketing research, Oxford, Blackwell.

Atkinson A., Riani C., Cerioli A., (2004), *Exploring Multivariate Dat with the Forward Search,* Springer, New York.

Atlas R., Overall J., (1994), *Comparative Evaluation of Two Superior Stopping Rules for Hierarchical Cluster Analysis,* Psychometrika, 59.

Bajgier S., Aggarwal L., (1991), *Powers of goodness-of-fit tests in detecting balanced mixed normal distributions,* Educational and Psychological Measurement, 51.

Benzecri J. (1973), *L'Anayse des donees. Tome 2: L'Analyse des correspondances*, Dunod.

Blashfield R., (1976), *Mixture Model Tests of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods,* Psychological Bulletin, 83.

Brusco M, Stahl M., (2004), *Variable Selection for Cluster Analysis* , Springer New York;

Brusco M., Cradit D., (2001), *A Variable-Selection Heuristic for K-Means Clustering*, Psychometrika, 66.

Carmone F. 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, No. 4

Chang W., (1983), *On using principal components before separating a mixture of two multivariate normal distributions,* Applied Statistics, 32.

Cormack R., (1971), *A review of classification (with discussion),* Journal of Royal Statistical Society, series A, 134.

Dash M, Liu H., (2000) *Feature selection for clustering* , Proceedings of Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, (PAKDD).

Denoeud L., Garreta H., Guénoche A. (2005), *Comparison of distance indices between partition*, Conference „International Symposium on Applied Stochastic Models and Data Analysis" (ASMDA2005).

De Soete G. (1986), *Optimal variable weighting for ultrametric and additive tree clustering*, Quality & Quantity, 20.

De Soete G. (1988), *OVWTRE: a program for optimal variable weighting for ultrametric and additive tree fitting*, Journal of Classification, 5.

De Soete G., Carroll J., (1994), *K-means clustering in a low-dimensional Euclidean space,* New approaches in classification and data analysis, Berlin, Springer.

DeSarbo W.S., Carroll J.D., Clark L.A., Green P.E. (1984), *Synthesized clustering: a method for amalgamating clustering bases with differential weighting of variables*, Psychometrika, vol. 49, no. 1, 57-78.

Devaney M. , Ram A., (1997), *Efficient feature selection in conceptual clustering*, Proceedings of the Fourteenth International Conference on Machine Learning, Nashville.

Dietterich, T. G., (1998). *Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms*, Neural Computation, 10.

Domański Cz., (1979), *Statystyczne Testy Nieparametryczne,* PWE Warszawa.

Domański Cz., Pruska K., (2007), *Nieklasyczne metody statystyczne,* Polskie Wydawnictwo Ekonomiczne, Warszawa.

Donoghue J.R. (1995), *Univariate screening measures for cluster analysis*, Multivariate Behavioral Research, 30.

Duda R., Hart P., (1973), *Pattern Classification and Scene Analysis,* Wiley, New York.

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

Everitt B., (2001), *Cluster Analysis,* HEB, London.

Fisher D., (1987), *Knowledge acquisition via incremental conceptual clustering,* PhD. Thesis, University of California, Irvine.

Fletcher J., Satz P., (1985), *Cluster Analysis and the search for learning disabilities subtypes,* w red. Rourke B., *Neuropsychology of learning disabilities: Essentilas of subtypes analysis.* New York, Guilford.

Fowlkes E.B., Gnanadesikan R., Kettenring J.R. (1987), *Variable selection in clustering and other contexts*, W: C.L. Mallows (red.), Design, data, and analysis, Wiley, New York.

Fowlkes E.B., Gnanadesikan R., Kettenring J.R. (1988), *Variable selection in clustering*, Journal of Classification 5.

Fraiman R., Justel A., Svarc M., (2008), *Selection of Variables for Cluster Analysis and Classification Rules,* JASA, 103.

Friedman J., Meulman J. (2004), *Clustering Objects on Subsets of Attributes*, Journal of the Royal Statistical Society, Series B 66.

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

Gatnar E., Walesiak M., (2004), *Metody Statystycznej Analizy Wielowymiarowej w Badaniach Marketingowych,* Wydawnictwo AE we Wrocławiu.

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

Gnanadesikan R., Kettenring J.R., Tsao S.L. (1995), *Weighting and selection of variables for cluster analysis,* Journal of Classification, 12.

Goeffe W., Ferrier G., Rogers J., (1994), *Global Optimization of Statistical Functions with Simulated Annealing,* Journal of Econometrics 60.

Gold E., Hoffman P., (1976), *Flange Detection Cluster Analysis,* Multivariate Behavioral Research, 11.

Grabiński T., (1992), *Metody taksonometrii,* Wydawnictwo AE, Kraków.

Green P., Carmone F., Kim J.,(1990), *A preliminary study of optimal variable weighting in k-means clustering,* Journal of Classification, 7.

Green P., Krieger A., (1995), *Alternative approaches to cluster-based market segmentation,* Journal of the Market Research Society, 37.

Guyon I., Elisseeff A. (2003), *An introduction to variable and feature selection*, Journal of Machine Learning Research, 3.

Hoff P., (2006), *Model-based subspace clustering,* Bayesian analysis 1.

Hubert L., Arabie P., (1985), *Comparing Partitions,* Journal of Classification 2.

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

Ichino M., Sklansky J., (1985), *The relative neighborhood graph for mixed feature variabless,* Pattern Recognition 18.

Jajuga K., (1989), *Podstawowe metody analizy wielowymiarowej w przypadku występowania zmiennych mierzonych na różnych skalach,* Praca w ramach CPBP, AE Wrocław.

Kaufman L., Rousseeuw P., (1990), *Finding Groups in Data,* John Wiley&Sons.

Kendall M., Buckland W., (1986), *Słownik terminów statystycznych,* PWE Warszawa.

Kleinberg J., (2002), *An Impossibility Theorem for Clustering.* Advances in Neural Information Processing Systems (NIPS).

Korzeniewski J., (2005), *Propozycja nowego algorytmu wyznaczającego liczbę skupień*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, Taksonomia 12.

Korzeniewski J., (2009), *An entropy based non-wrapper approach for choosing variables in cluster analysis*, Acta Universitatis Lodziensis, Folia Oeconomica, 172.

Korzeniewski J., (2010), *Badanie odporności metody HINoV na błędnie zadaną liczbę skupień w zbiorze danych,* Taksonomia 17: Klasyfikacja i analiza danych – teoria i zastosowania, Nr 107, Wrocław.

Korzeniewski J., (2012), *Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień,* Taksonomia 19, Klasyfikacja i analiza danych – teoria i zastosowania, WUE Wrocław.

Kruskal W., (1972), *Linear transformation of multivariate data to reveal clustering*, *in* Shepard R., Romney A. and Nerlove S., (eds), *Multidimensional scaling: Theory and Applications in the Behavioural Sciences*, Vol. 1, Seminar Press, London.

Kuiper F., Fisher L., (1975), *A Monte Carlo Comparison for Six Clustering Procedures,*. Biometrics, 31.

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

Makarenkov V., Legendre P. (2001), *Optimal variable weighting for ultrametric and additive trees and K-means partitioning methods and software*, Journal of Classification, 18.

Martinez W., Martinez A., (2001) , *Computational statistics handbook with MATLAB* . Boca Raton: Chapman&Hall.

McIntyre R., Blashfield R., (1980), *A Nearest-Centroid Technique for Evaluating the Minimum Variance Clustering Procedure,* Multivariate Behavioral Research, 15.

Migdał-Najman K., Najman K. (2005), *Analityczne metody ustalania liczby skupień*, Taksonomia 12: Klasyfikacja i analiza danych – teoria i zastosowania, Nr 1076, Wrocław.

Mikulec A., (2010), *Zastosowanie metod statystycznych do oceny efektywności i klasyfikacji systemów emerytalnych,* nieopublikowana praca doktorska na Wydziale Ekonomiczno-Socjologicznym Uniwersytetu Łódzkiego.

Milligan G., **(**1980), *An examination of the effect of six types of error perturbation on fifteen clustering algorithms* , Psychometrika, 45.

Milligan G.**, (**1985), *An Algorithm for Generating Artificial Test Clusters* , Psychometrika, 50.

Milligan G., Cooper M., (1987), *Methodological review: Clustering methods*, Applied Psychological Measurement, 11.

Milligan G., Cooper M., (1988), *A study of variable standarization*, Journal of Classification, 5.

Milligan G., (1989), *A validation study of a variable weighting algorithm for cluster analysis*, Journal of Classification, 6.

Milligan G., (1994), *Issues in applied classification: selection of variables to cluster*, Classification Society of North America Newsletter, 37.

Milligan G., (1996), *Clustering Validation: Results and Implications for Applied Analyses,* [w] Clustering and Classification, red, P. Arabie, L. Hubert, G. de Soete, World Scientific, Singapore .

Mojena R. (1977), *Hierarchical grouping methods and stopping rules: an evaluation*, Computer Journal 1977, Vol. 20, No 4.

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

Najman K., Migdał-Najman K., (2006), *Wykorzystanie indeksu Silhouette do ustalania optymalnej liczby skupień,* Wiadomości Statystyczne, 6.

Nowak E. (1985), *Wskaźnik podobieństwa wyników podziałów*, Przegląd Statystyczny, 1.

Pawełek B., (2008), *Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie.

Pearson K., (1894), *Contributions to the theory of mathematical evolution,* Philosophical Transactions of the Royal Society of London, series A, 185.

Pełka M., Wilk J., (2010), *Metody selekcji zmiennych symbolicznych w zagadnieniach klasyfikacji,* Klasyfikacja i analiza danych – teoria i zastosowania, Taksonomia 17, Prace Naukowe UE we Wrocławiu nr 107.

Pociecha J. (1982), *Kryteria oceny procedur taksonomicznych*, Przegląd statystyczny, Nr 1/2.

Press W., Flannery B., Teukolsky S., Vetterling W., (1986) *Numerical Recipes , The Art of Scientific Computing,* Cambridge, Cambridge University Press.

Price L., (1993), *Identifying Cluster Overlap with NORMIX Population Membership Probabilities,* Multivariate Behavioral Research, 28.

Qiu W., Joe H., (2006), *Generation of random clusters with specified degree of separation,* Journal of Classification 23.

Raftery, A.E. and Dean, N. (2006). *Variable Selection for Model-Based Clustering,* Journal of the American Statistical Assocation, 101.

Ratkowsky, Lance, (1978), *A criterion for determining the number of groups in a classification,* Australian Computer Journal, 10.

Schaffer C., Green P., (1998), *Cluster-based market segmentation: Some further comparisons of alternative approaches, ,* Journal of the Market Research Society, 40.

Sokołowski A. (1992), *Empiryczne testy istotności w taksonomii*, Zeszyty Naukowe AE w Krakowie, Seria specjalna: Monografie nr 108.

Steinley D., (2004), *Standardising variables in K-means clustering* , [w] Banks D., House L., McMorris F., Arabie P., Gaul W. (red.), Classification, Clustering and Data Mining Applications, Springer, New-York.

Steinley D., (2006), *K-means clustering, A half-century synthesis* , British Journal of Mathematical and Statistical Psychology, 59.

Steinley D., Brusco M. (2007a), *A New Variable Weighting and Selection Procedure for K-means Cluster Analysis* , Multivariate Behavioral Research.

Steinley D., Brusco M. (2007b), *Initializing k-means batch clustering: A critical evaluation of several techniques,* Journal of Classification 24.

Steinley D., Brusco M. (2008), *Selection of Variables in Cluster Analysis: An Empirical Comparison of Eight Procedures,* Psychometrika 73.

Steinley D., Henson R., (2005), *OCLUS: An Analytic Method for Generating Clusters with Known Overlap* , Journal of Classification 22.

Stoddard M., (1979), *Standarisation of measures prior to cluster analysis,* Biometrics 35.

Tadesse M., Sha N., Vanucci M., (2005), *Bayesian Variable Selection in Clustering High-Dimensional Data,* Journal of the American Statistical Association v. 100.

Tadesse M., Kim S., Vanucci M., (2006), *Variable selection in clustering via Dirichlet process mixture models,* Biometrika 93.

Talavera L., (2000), *Dependency-Based Feature Selection for Clustering Symbolic Data* , Intelligent Data Analysis 4.

Tibshrani R., Walther G., Hastie T., (2001), *Estimating the Number of Clusters in a Datase via the Gap Statistic,* Journal of the Royal Statistical Society 32.

Torgerson W., (1952), *Multidimensional Scaling: I.Theory and method*, Psychometrika 17.

van Buuren S., Heiser W., (1989) *Clustering N Objects into K Groups Under Optimal Scaling of Variables,* Psychometrika 54.

Walesiak M., (1993), *Strategie postępowania w badaniach statystycznych w przypadku zbioru zmiennych mierzonych na skalach różnego typu,* Badania Operacyjne i Decyzje 1.

Walesiak M., (2002), *Uogólniona miara odległości w statystycznej analizie wielowymiarowej,* Wydawnictwo AE Wrocław.

Walesiak M., (2005), Problemy selekcji i ważenia zmiennych w zagadnieniu klasyfikacji, W: K. Jajuga, M. Walesiak, Klasyfikacja i analiza danych

– teoria i zastosowania, Taksonomia 12, Prace Naukowe AE we Wrocławiu nr 1076.

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

Walesiak M., Dudek A., (2008), *Identification of noisy variables for nonmetric and symbolic data in cluster analysis*, W: C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (red.), Data analysis, machine learning and applications, Springer-Verlag, Berlin, Heidelberg.

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

Waller N., Underhill J, Kaiser H., (1999), *A Method for Generating Simulated Plasmodes and Artificial Test Clusters with User-Defined Shape, Size and Orientation.* Multivariate Behavioral Research, 34.

Wieczorkowski R., Zieliński R., *Komputerowe generatory liczb losowych,* Wydawnictwa Naukowo Techniczne, Warszawa 1997.

Witten D., Tibshirani R., (2010), *A framework for feature selection in clustering,* Journal of the American Statistical Association 105.

Zhou J., He X., (2008), *Dimension Reduction Based on Canonical Correlation and Variable Filtering,* Annals of Statistics 36.

Załacznik 1. Oznaczenia i symbole

n – liczba obserwacji/obiektów w zbiorze danych;
K –liczba skupień, k – indeks jednego skupienia;
C_k - zbiór obserwacji z k-tego skupienia;
$|A|$, card(A) - moc zbioru A;
n_k - liczebność k-ego skupienia, $n_k = |C_k|$;
$\boldsymbol{X}$ – macierz danych o wymiarach $n \times V$
x_{iv} - element macierzy $\boldsymbol{X}$, czyli wartość i-ego obiektu na v-ej zmiennej;
d_{ij} - odległość pomiędzy i-tym a j-tym obiektem;
$\boldsymbol{D}$ - macierz odległości d_{ij};
α_k - proporcjonalny udział k-ego skupienia wśród wszystkich obiektów, $\sum_{k=1}^{K} \alpha_k = 1$;
θ_k - wektor parametrów charakteryzujący k-te skupienie;
Σ_k - macierz kowariancji k-ego skupienia;
σ_{ij} - kowariancja pomiędzy i-tą a j-tą zmienną;
z_i - wektor n-wymiarowy przynależności i-tego obiektu do skupień; $z_{is} = 1$ gdy i-ty obiekt należy do skupienia o numerze s oraz $z_{is} = 0$ gdy i-ty obiekt nie należy do skupienia o numerze s;
c - funkcja określająca przynależność obiektów do skupień, $c(i) = s$ gdy i-ty obiekt należy do skupienia o numerze s
I_s - macierz jednostkowa stopnia s;
J_s - macierz jedynkowa stopnia s;
V - liczba zmiennych;
$\boldsymbol{V}$ - wektor wszystkich zmiennych, $\boldsymbol{V} = [v_1, \ldots, v_V]^T$;
T - liczba zmiennych istotnych dla struktury skupień;
M - liczba zmiennych nieistotnych dla struktury skupień, $V=T+M$;
VV - zbiór wszystkich zmiennych;
TT -zbiór zmiennych istotnych dla struktury skupień;
MM - zbiór zmiennych nieistotnych dla struktury skupień;
$\sigma^2(v)$ - wariancja v-ej zmiennej;

$r(v)$; r_j - obszar zmienności (rozstęp) zmiennej v ; obszar zmienności j-ej zmiennej;

PP - zbiór wszystkich podzbiorów zbioru wszystkich zmiennych z wyłączeniem podzbioru pustego;

P - podzbiór zbioru wszystkich zmiennych, $P \in PP$;

$PP(s)$ - zbiór wszystkich podzbiorów s elementowych zbioru wszystkich zmiennych, $PP(s) \subset PP$;

$\pi(P)$ - podział zbioru danych w oparciu o zmienne należące do zbioru *P;*

$\pi = \{C_1, C_{2,} \ldots, C_k\}$ - podział zbioru danych na K skupień $C_1, C_{2,} \ldots, C_k$.

Załacznik 2. Spis programów komputerowych

W trakcie pracy nad tą monografią powstały następujące programy komputerowe napisane przez autora w środowisku języka Delphi:

1. Program do generowania wielowymiarowych struktur skupień zmiennych ciągłych działający według algorytmu OCLUS.
2. Program do generowania struktur skupień zmiennych binarnych.
3. Program do metody *KO_ciągłe*.
4. Program do metody *KO_porządkowe*.
5. Program do metody *KO_binarne*.
6. Program do metody *KO_meta*.
7. Program do metody Ichino.
8. Program do metod opartych na entropii.
9. Program do metody uśredniania zmiennych.
10. Program do metody HINoV.
11. Program do metody VS-KM.
12. Program do metody VAF z indeksem skupialności.

Od redakcji

Dr Jerzy Korzeniewski jest adiunktem w Katedrze Metod Statystycznych Instytutu Statystyki i Demografii Uniwersytetu Łódzkiego. Jego zainteresowania naukowo-badawcze koncentrują się wokół analizy skupień, analizy obrazów, wcześniej odporności statystycznej. Dr Jerzy Korzeniewski jest autor około 30 artykułów naukowych oraz wielu opracowań w ramach projektów badawczych. Dr Jerzy Korzeniewski prezentował wyniki swoich badań na wielu konferencjach naukowych w kraju i za granicą, w tym w Singapurze, Niemczech, RPA, Anglii, Czechach i Portugalii.

W 1999 roku obronił rozprawę doktorską *Metody badania odporności testów statystycznych*, przygotowaną pod kierunkiem prof. Czesława Domańskiego. W latach 1991–1992 był na sześciomiesięcznym stypendium naukowym na University of Swansea, w Walii. W roku 1993 był na czteromiesięcznym Business Course na University of Nijmegen, w Holandii. W latach 1994–1997 współuczestniczył w realizacji międzynarodowego programu COST zatytułowanego: Unemployment During Economic Transition, a w latach 2010–2011 był kierownikiem i głównym wykonawcą projektu KBN na temat: *Metody wybierania zmiennych w analizie skupień i ich zastosowania*. Od wielu lat prowadzi wykłady specjalistyczne na kierunku „Informatyka i Ekonometria" na Wydziale Ekonomiczno-Socjologicznym oraz wykłady i ćwiczenia z kilku przedmiotów o charakterze matematycznym, statystycznym jak również z teorii ryzyka i matematyki finansowej. Ponadto w ramach programu Erasmus prowadzi zajęcia z tych przedmiotów dla studentów obcokrajowców.

Niniejsza monografia stanowi podsumowanie badań i doświadczeń autora dotyczących metodologii selekcji zmiennych w analizie skupień. Zaproponowane nowe procedury selekcji zmiennych są efektywniejsze od dotychczas opracowanych a ponadto elastyczniejsze pod względem możliwości aplikacyjnych. Opracowane metody stanowią istotny wkład do teorii selekcji zmiennych w analizie skupień.

Summary

Methods of Variable Selection in Cluster Analysis. New procedures

Cluster analysis also called unsupervised learning is devoted to the classification of data set objects into disjoint clusters or classes homogenous inside each class and heterogenic between the classes. Properly carried out it allows to divide the data set into groups in order to understand better the information contained within them. Cluster analysis is very important in many sciences as well in the analysis of socio-economic phenomena. The tasks of cluster analysis cannot be achieved without the proper selection of variables which create a cluster structure.

The main objective of this book is to assess the methods of variable selection developed by others and to propose original new procedures as well as the possibilities of their application to socio-economic types of data.

The first chapter is a general introduction to the problems and measures of cluster analysis, description of the objectives and assumptions of variable selection task. It also contains the description of generated data sets.

The second chapter is devoted to the model approaches to variable selection in cluster analysis.

The third chapter is devoted to heuristic approaches to the task of variable selection. Heuristic approaches constitute the bulk of the volume of all methods constructed so far and the best of them are superior to the model approaches. Among the best heuristic procedures one should mention HINoV, VS-KM, VAF with clusterability index. The best methods were once again thoroughly tested by the author on the same data sets on which new proposals were tested in order to create a solid foundation for the sake of fair comparison.

The fourth chapter contains the proposal of a new method of variable selection which is based on the linear coefficient between the distances between two objects measured on different set of variables. In this way we get a very universal tool of variable selection which can be applied to both strong and weak measurement scales. To some extent we also skip the necessity of the knowledge of the number of clusters in a data set.

The fifth chapter is entirely devoted to the applicational side of the methods presented in the book to the analysis of real world data sets, especially the ones of the socio-economic nature. All the methods which were considered best so far are used in this chapter and compared with the author's proposals.