

WOLFGANG HARTMANN
Dresden

Współczesne metody analizy danych w socjologii*

(Praca wpłynęła do Redakcji 15.07.1982)

Socjolog, stosujący w swoich pracach metody empiryczne, podobnie jak badacze w zakresie innych nauk społecznych posługujący się tymi metodami, staje wobec problemu interpretowalności zebranych danych. Semantyczne przesłanki dla tej interpretacji uwarunkowane są tym, w jaki sposób hipotezy wysnute z rozważań teoretycznych przełożone zostaną na różne wskaźniki i reakcje badanych jednostek. Złożoność sytuacji społecznych, które przy różnych spojrzeniach badacza wyjawiają różne swoje strony, w znacznym stopniu wyklucza możliwość jednoznacznego przyporządkowania badanym cechom ściśle określonych treści teoretycznych.

W ostatnich latach socjologowie i psychologowie stosujący metody matematyczne dużo uwagi poświęcali wykryciu i mierzeniu ukrytych postaw i zachowań, typowych dla różnych grup społecznych. Matematyczna analiza danych, pochodzących z ankiet i testów, stwarza metodologicznie nowe problemy, przede wszystkim z tego powodu, że pierwotne informacje mają tu raczej jakościowy, niż ilościowy charakter. Ponadto takie założone pojęcia jak inteligencja lub aktywność społeczna stanowią tak szeroki wachlarz różnorodności, że mierzenie ich za pomocą konwencjonalnych jednowymiarowych skal okazuje się niewystarczające. W tym przy-

* Tłumaczył Ryszard Zieliński.

padku zagadnienie polega nie tylko na tym, żeby na poszczególnych skalach (w poszczególnych "wymiarach") wybrać odpowiednie jednostki, ale także na tym, żeby wyjaśnić znaczenie poszczególnych skal, przy czym dobrać je tak, żeby były możliwie od siebie niezależne i żeby dostatecznie dokładnie określały te cechy, które w ogólnym przypadku nie są bezpośrednio mierzalne.

W niniejszym artykule przedstawimy najpierw różne metody skalowania wielowymiarowego. Ta technika wielowymiarowej analizy danych jest szczególnie przydatna w różnych metodach analizy czynnikowej i podobnie jak analiza czynnikowa może być traktowana jako metoda regresji nieliniowej z określoną funkcją, której parametry należy oszacować. Nie będziemy tu jednak zagłębiać się w matematyczne metody estymacji, których do tej pory dużo skonstruowano, ale główną uwagę skupimy na ogólnym sformułowaniu problematyki skalowania wielowymiarowego jako pewnej techniki analizy danych.

W psychologii szczególnie ważną rolę odgrywa problematyka skalowania międzyosobniczych i wewnątrzosobniczych różnic w postawach, ocenach i zachowaniach; w ostatnich latach dla oceny tych różnic stosowano z powodzeniem pewne proste rozszerzenie metod skalowania wielowymiarowego, a mianowicie tak zwane metody skalowania wielowymiarowego różnic indywidualnych. W szczególności podjęto przy tym próby interpretowania trójwymiarowych i wielowymiarowych macierzy odległości subiektywnych jako macierzy odpowiednio ważonych odległości euklidesowych. W pierwszych dwóch rozdziałach analizujemy za pomocą metody skalowania wielowymiarowego i metody skalowania wielowymiarowego różnic indywidualnych pewien praktyczny przykład, a w trzecim rozdziale - dla porównania - przedstawiamy wyniki analizy tego samego przykładu za pomocą hierarchicznej analizy skupień.

1. METODY SKALOWANIA WIELOWYMIAROWEGO

Celem skalowania wielowymiarowego jest przyporządkowanie danym n badanym obiektom (wskaźnikom, bodźcom) A_1, \dots, A_n punktów X_1, \dots, X_n w r -wymiarowej przestrzeni metrycznej w taki sposób

żeby podobieństwa lub różnice pomiędzy obiektami reprezentowane były możliwie dobrze - przy pewnym ustalonym kryterium tej dobroci - przez odległości pomiędzy punktami.

Każdemu z n badanych obiektów A_i , $i=1, \dots, n$, przyporządkowane zostaje więc r liczb x_{ik} , $k=1, \dots, r$, opisujących położenie punktu X_i w r -wymiarowym prostokątnym układzie współrzędnych. Położenie zbioru punktów $\{X_1, \dots, X_n\}$ zostaje przy tym opisane za pomocą tzw. macierzy współrzędnych $X = (x_{ik})$ o wymiarach $n \times r$; będziemy tu również używali nazwy konfiguracja punktów.

Przy ustalonej ("obiektywnej") metryce d można macierzy współrzędnych X przyporządkować macierz odległości $D = (d_{ij})$, $i, j = 1, \dots, n$, taką, że d_{ij} jest odległością punktów X_i, X_j . Najbardziej znaną jest z pewnością metryka euklidesowa

$$(1.1) \quad d_{ij} = \left(\sum_{k=1}^r (x_{ik} - x_{jk})^2 \right)^{1/2},$$

ale poza tym do być może najbardziej znanych przykładów należą:

(a) rodzina metryk Minkowskiego

$$(1.2) \quad d_{ij} = \left(\sum_{k=1}^r w_k |x_{ik} - x_{jk}|^q \right)^{1/q}$$

z parametrami $w_k \geq 0$, $q \geq 1$,

(b) rodzina metryk wykładniczych

$$(1.3) \quad d_{ij} = \log_q \left(\sum_{k=1}^r q^{w_k |x_{ik} - x_{jk}|} \right)$$

z parametrami $w_k > 0$, $q > 1$.

Metryka euklidesowa jest szczególnym przypadkiem metryki Minkowskiego dla $q = 2$ oraz $w_k \equiv 1$. W przypadku jednowymiarowym ($r = 1$), gdy zrezygnujemy z wprowadzania wag w_k , wszystkie te metryki przyjmują postać $d_{ij} = |x_{i1} - x_{j1}|$.

Punktem wyjścia dla przyporządkowania n badanym obiektom A_i punktów X_i jest w ogólnym przypadku $(n \times n)$ -wymiarowa macierz S tak zwanych "odległości subiektywnych" (podobieństw, różnic) s_{ij} dla par (A_i, A_j) obiektów. Wiersze i kolumny macierzy S odpowiadają kolejnym obiektom A_1, \dots, A_n , a element s_{ij} wyraża w pewien sposób empirycznie oszacowany stopień zróżnicowania obiektów A_i, A_j . Odległości subiektywne s_{ij} otrzymuje się albo na drodze porównania każdej pary (A_i, A_j) elementów przez ekspertów, przez badane osoby lub przez ankietowanych i bezpośrednie przypisanie przez nich każdej parze odpowiedniej liczby w wielopunktowej skali porządkowej, albo pośrednio z innych danych. Szacowanie stopnia podobieństwa przez takie bezpośrednie porównania występuje przede wszystkim w psychologii, a szczególnie w psychofizyce; w tym ostatnim przypadku istnieją możliwości badania różnic pomiędzy ocenami subiektywnymi a wynikami obiektywnych pomiarów odpowiednich wielkości fizycznych (np. zróżnicowanie kolorów, dźwięków, obrazów geometrycznych). W socjologii do tej pory analizowano najczęściej macierze S , otrzymane za pośrednictwem innych danych empirycznych; były to zwykle macierze kowariancji, macierze korelacji, macierze wskaźników asocjacji lub innych tego typu wskaźników.

Do skalowania wielowymiarowego nadają się jednak również inne rodzaje empirycznie dostępnych informacji. Można na przykład za miarę podobieństwa dwóch obiektów A_i, A_j (np. dwóch wiadomości przekazanych alfabetem Morse'a) przyjąć prawdopodobieństwo zdarzenia polegającego na tym, że jeden z tych obiektów uznany zostanie za drugi. Podobnie za subiektywną miarę odległości pomiędzy dwiema określonymi grupami społecznymi można na przykład przyjąć liczbę kontaktów zawodowych pomiędzy członkami tych grup lub liczbę par połączonych węzłami sympatii. Za miarę odległości dwóch gatunków biologicznych można przyjąć liczbę kontaktów seksualnych pomiędzy osobnikami różnych gatunków. Stopień podobieństwa różnych pojęć abstrakcyjnych można np. mierzyć w ten sposób, że ankietowana osoba możliwie bez namysłu na podane jej pojęcie odpowiada pojęciem, które jej z danym pojęciem najbardziej się kojarzy; w takiej sytuacji nie

zawsze kojarzone są "bliskie" pojęcia - np. "ogień" kojarzy się mocno z "wodą", chociaż oba te obiekty są mało podobne do siebie.

Macierze S odległości subiektywnych s_{ij} , które mają być analizowane metodą skalowania wielowymiarowego, powinny - z dokładnością do błędów losowych - mieć następujące własności (co czasami zagwarantowane jest przez samą naturę problemu):

- (a) odległość subiektywna danego obiektu A_i od siebie samego powinna być możliwie bliska zeru, a przynajmniej mniejsza od subiektywnej odległości A_i od każdego innego obiektu,
- (b) subiektywna odległość obiektu A_i od obiektu A_j powinna być mniej więcej równa subiektywnej odległości A_j od A_i .

Macierz S powinna więc, przynajmniej mniej więcej, być macierzą symetryczną o zerowej głównej przekątnej. Często obie te własności zagwarantowane są przez samą metodę szacowania odległości subiektywnych, istnieją jednak sytuacje, w których jest to nawet teoretycznie niemożliwe. Przykładem takiej sytuacji jest badanie związków sympatii pomiędzy członkami danej grupy: związki takie nie są na ogół symetryczne, lecz zależą od "kierunku". Analiza subiektywnych "odległości" tego typu wymaga innych metod niż omawiana tu metoda skalowania wielowymiarowego.

Metoda skalowania wielowymiarowego polega na tym, żeby przy ustalonej metryce d oraz przy ustalonym wymiarze r wyznaczyć taką konfigurację punktów X , przy której odległości obiektów (macierz D) i odległości subiektywne (macierz S) będą różniły się od siebie możliwie mało (w określonym sensie). Ze względu na postać kryterium zgodności tych macierzy, rozróżnia się metryczne i niemetryczne ("porządkowe") metody skalowania wielowymiarowego.

W metodach metrycznych dąży się do wyboru konfiguracji zapewniających bezpośrednią zgodność odpowiednich odległości subiektywnych i obiektów. Poszukuje się na przykład macierzy $X = (x_{ik})$ minimalizującej sumę kwadratów różnic

$$(1.4) \quad \sum_{i,j} (d(X_i, X_j) - s_{ij})^2.$$

To kryterium, przy danej macierzy odległości subiektywnych S , nie wyznacza rozwiązania w sposób jednoznaczny; przy danej macierzy d można bowiem konfigurację X poddawać pewnym przekształceniom, nie zmieniając przy tym macierzy D . Jeżeli d jest metryką euklidesową, to takimi przekształceniami są wszelkie przesunięcia i obroty. Wynika stąd, że jeżeli istnieje optymalna konfiguracja \hat{X} minimalizująca (1.4), to istnieje nieskończenie wiele różnych konfiguracji tak samo optymalnych. Metody rozwiązywania problemu polegają tu na wyznaczeniu najpierw pewnego rozwiązania \hat{X} , a następnie poddaniu go takim przekształceniom (niezmienniczym ze względu na wartość kryterium (1.4)), które prowadzą do konfiguracji nadającej się do wygodnej interpretacji.

W metodach niemetrycznych dąży się tylko do uzyskania zgodności pomiędzy "rangami" odpowiednich odległości w tym sensie, że nierówność $s_{ij} < s_{kl}$ odległości subiektywnych powinna w zasadzie pociągnąć za sobą odpowiednią nierówność $d_{ij} < d_{kl}$ odległości obiektywnych. Metody tego rodzaju mają tę zaletę, że dla uzyskania konfiguracji punktów X potrzebne są tylko rangi odległości subiektywnych. Formalnie mówiąc, konfiguracja X powinna być niezmiennicza ze względu na dowolne ściśle rosnące przekształcenia t danych s_{ij} i powinna być rozwiązaniem zadania minimalizacji kryterium postaci

$$(1.5) \quad \sum_{i,j} (d(X_i, X_j) - t(s_{ij}))^2$$

względem konfiguracji X i ściśle rosnących funkcji t . W celu uniknięcia sztucznego obniżenia wartości tego kryterium przez "skurczenie" konfiguracji, potrzebna jest tu odpowiednia normalizacja. Zwykle więc zamiast (1.5) stosuje się jedno z kryteriów

$$(1.6) \quad \frac{\sum_{i,j} (d_{ij} - t(s_{ij}))^2}{\sum_{i,j} d_{ij}^2} \quad \text{lub} \quad \frac{\sum_{i,j} (d_{ij} - t(s_{ij}))^2}{\sum_{i,j} (d_{ij} - d_{..})^2},$$

gdzie d_{ij} jest średnią arytmetyczną wartości d_{ij} , $i < j$. Zadanie minimalizacji (1.6) również nie ma jednoznacznego rozwiązania. Jeżeli istnieje optymalne rozwiązanie \hat{X} tego problemu, to optymalne są również różne konfiguracje, które można otrzymać z konfiguracji \hat{X} na drodze przekształceń nie zmieniających rang odległości d_{ij} . Przy ustalonej metryce d , niemetryczne zagadnienie z kryterium (1.6) jest więc jeszcze bardziej nieokreślone niż odpowiednie zagadnienie metryczne. Do tego problemu powrócimy jeszcze później.

Z matematycznego punktu widzenia, problematyka skalowania wielowymiarowego jest problematyką estymacji parametrów regresji nieliniowej. W przypadku metrycznym stosuje się tu przede wszystkim metody numeryczne algebry liniowej, a w przypadku niemetrycznym do tej pory najczęściej posługiwano się numerycznymi iteracyjnymi metodami optymalizacji nieliniowej w połączeniu z metodami tak zwanej regresji izotonicznej (por. Barlow i in. [1]). Do najbardziej znanych metod metrycznych skalowania wielowymiarowego należy metoda Torgersona [17], a klasyczne metody niemetryczne pochodzą od Kruskala [8], [9]. Obie metody były wielokrotnie modyfikowane i ulepszone. Szczegółowy opis odpowiednich algorytmów można znaleźć w monografii autora [5].

Współczesne metody skalowania wielowymiarowego nie ograniczają się wyłącznie do wyznaczania konfiguracji X przy ustalonym wymiarze r i przy ustalonej metryce d . Oczywiście jest, że konfiguracje X obliczone przy większych wartościach r dają lepszą zgodność odległości subiektywnych i obiektywnych niż przy mniejszych wartościach tego parametru. Dowodzi się na przykład (patrz [10]), że jeżeli $r = n-1$ i jeżeli macierz odległości empirycznych $S = (S_{ij})$ spełnia odpowiednie warunki, to można uzyskać idealną zgodność, przy której suma kwadratów (1.5) przyjmuje wartość zero. Konfiguracje X przy wysokich wymiarach mają jednak tę wadę, że są trudne do interpretowania i że obok istotnych informacji zawartych w macierzy S odzwierciedlają również zawarte w nich błędy przypadkowe. Z tych powodów wymiar r powinien być tak mały, jak to jest tylko możliwe (eli-

minacja błędów przypadkowych!), ale z drugiej strony powinien być tak duży, jak to jest potrzebne (uwzględnienie informacji istotnych!). Zagadnienia określenia wymiaru r "tkwiącego" w danej macierzy danych S jest więc raczej zagadnieniem statystycznym niż numerycznym. Jest to zagadnienie nie w pełni rozwiązane; przypomina ono zagadnienie oszacowania liczby istotnych czynników w analizie czynnikowej oraz zagadnienie oszacowania liczby istotnych zmiennych w analizie regresji.

Poza wyznaczeniem optymalnej konfiguracji X i optymalnego wymiaru r , od metody skalowania wielowymiarowego można wymagać, aby w optymalny sposób (w sensie możliwie dobrej zgodności macierzy danych S i macierzy odległości D) wyznaczała parametry opisujące daną rodzinę metryk (np. parametry q i w_k , $k = 1, \dots, r$, w (1.2) i (1.3)). Wybór optymalnej metryki d jest istotny na przykład wtedy, gdybyśmy chcieli uzyskać wskazówki co do "osobistej" metryki badanej osoby. Psycholodzy odkryli na przykład, że zmęczenie badanej osoby i jej mniejsza zdolność do koncentracji związane są ze wzrostem parametru q metryki Minkowskiego (1.2); wtedy odległość d_{ij} pomiędzy dwoma obiektami A_i oraz A_j zależy przede wszystkim od różnicy pomiędzy najbardziej różniącymi się między sobą współrzędnymi punktów X_i , X_j . W przypadku tzw. metryki "taksówkowej" (metryka Minkowskiego dla $q = 1$, $w_k \equiv 1$, czyli tzw. "city-block distance")

$$d_{ij} = \sum_{k=1}^r |x_{ik} - x_{jk}|$$

wszystkie różnice $|x_{ik} - x_{jk}|$, $k=1, \dots, r$, uwzględnione są w jednakowym stopniu, a w przypadku metryki Minkowskiego z $q = \infty$, $w_k \equiv 1$, mamy

$$d_{ij} = \max_{1 \leq k \leq r} |x_{ik} - x_{jk}|$$

i odległość d_{ij} jest determinowana tylko przez największą różnicę pomiędzy odpowiednimi współrzędnymi.

Ważnym zagadnieniem dla interpretacji wyników analizy przeprowadzonej metodą skalowania wielowymiarowego jest zagadnienie niezmienniczości konfiguracji X . W ogólnym przypadku punkty X_i konfiguracji można poddawać różnym przekształceniom bez zmiany jakości rozwiązania, tzn. bez zmiany tego, w jakim stopniu odległości subiektywne i obiektywne zostały do siebie dopasowane. W przypadku gdy d jest metryką euklidesową, wartości d_{ij} nie ulegają zmianie

(a) przy dowolnych przesunięciach i

(b) przy dowolnych obrotach i odbiciach układu współrzędnych.

W przypadku metody niemetrycznej, w której odgrywają rolę tylko rangi odległości, konfiguracja X może być poddawana ponadto wszelkim przekształceniom, zachowującym porządek. Na przykład pomnożenie jednej współrzędnej x_{ik} przez pewną stałą K powoduje $|K|$ -krotną zmianę odległości Minkowskiego, a więc konfiguracje otrzymane metodami niemetrycznymi można dowolnie "ściągać" lub "rozciągać". W celu zrekompensowania opisanych wyżej niezmienniczości rozwiązania, wprowadza się w ogólnym przypadku dodatkowe warunki normalizacji. W przypadku metod niemetrycznych z metryką euklidesową, można żądać, aby

(a) punkt ciężkości konfiguracji znajdował się w początku układu współrzędnych (translacja),

(b) tzw. główne osie konfiguracji pokrywały się z osiami układu współrzędnych (obroty),

(c) wariacje poszczególnych współrzędnych były równe jedności (jednokładność).

W przypadku ogólnej metryki Minkowskiego, niezmienniczość względem dowolnych obrotów musi być zastąpiona niezmienniczością względem dowolnych permutacji osi współrzędnych, a warunek (b) zostaje zastąpiony warunkiem uporządkowania kolejności współrzędnych według malejącej wariancji. W przypadku metody metrycznej, postulat (c) musi być oczywiście odrzucony. Z punktu widzenia praktycznych zastosowań często nie wygodnie jest interpretować konfiguracje otrzymane przy warunkach normalizacji narzuconych wyłącznie ze względu na prostotę matematyczną. Korzystając z pewnych dodatkowych, odpowiednio zebranych informa-

cji (np. wartości pewnych cech), można konstruować konfiguracje dogodniejsze do semantycznych interpretacji. Jest to takie samo zagadnienie, z jakim spotykamy się w analizie czynnikowej, gdzie ostatnio coraz częściej odchodzi się od rotacji uzasadnionych tylko czysto matematyczną prostotą odpowiednich struktur.

Przedstawiony niżej przykład zastosowania metody skalowania wielowymiarowego uwzględnia nie tyle sprawę pomiaru odpowiednich cech, ile i to przede wszystkim, te własności metody, które są związane z pewną syntezą danych. Oto ten przykład.

W NRD przeprowadzono szeroko zakrojone badania (około 10000 ankietowanych osób), mające na celu wykrycie struktur społecznych w populacji zatrudnionych w przemyśle państwowym. Każda z ankietowanych osób miała zaznaczyć na załączonej liście te zajęcia i hobby, które uprawia w wolnym czasie:

1. wykonuję prace typu rzemieślniczego w domu i w ogródku,
2. pielęgnuję swój pojazd,
3. uprawiam sport,
4. uczęszczam na imprezy sportowe,
5. uczęszczam na wykłady i kursy dokształcające,
6. kolekcjonuję znaczki pocztowe, monety, ... ,
7. uczęszczam na imprezy rozrywkowe, wieczorki taneczne,
8. chodzę do kina,
9. chodzę do teatru, na koncerty, wystawy,
10. sam uprawiam sztukę,
11. majsterkuję,
12. wykonuję robótki ręczne,
13. spotykam się z przyjaciółmi,
14. chodzę na piwo,
15. czytam książki, słucham radia lub muzyki z płyt,
16. oglądam telewizję;
17. gram w karty, szachy, itp.,
18. chodzę na spacer, wędrówki, wycieczki,
19. spędzam wolny czas z dziećmi,
20. właściwie to nie wiem, co robić z wolnym czasem,
21. mam inne zajęcia i zainteresowania.

Za podstawę wskaźnika s'_{ij} podobieństwa pomiędzy dwoma sposobami spędzania wolnego czasu przyjęto liczbę ankietowanych t_{ij} , którzy zaznaczyli zarówno czynność A_i , jak i czynność A_j . Spośród wielu stosowanych wskaźników podobieństwa wybrano, ze względu na jego szczególne własności metryczne, wskaźnik Tanimoto

$$s'_{ij} = \frac{t_{ij}}{t_{i.} + t_{.j} - t_{ij}}, \quad i, j = 1, \dots, n,$$

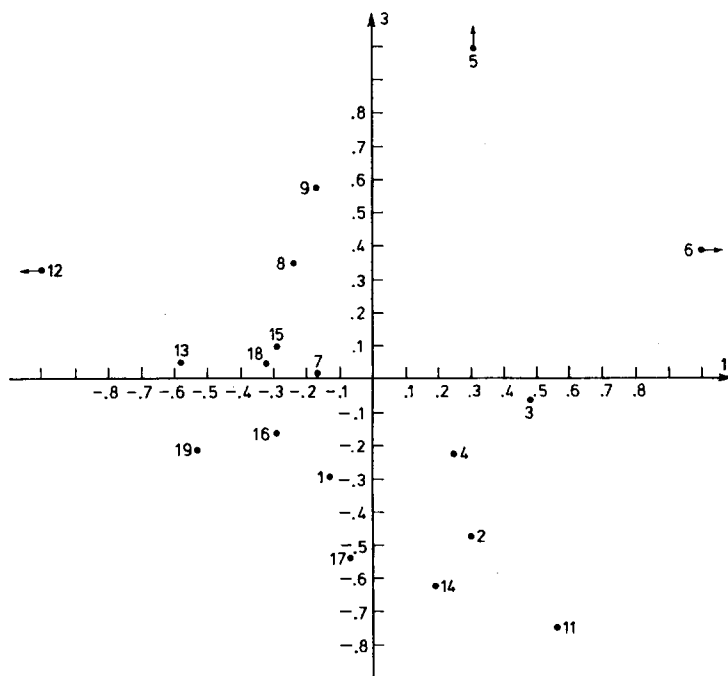
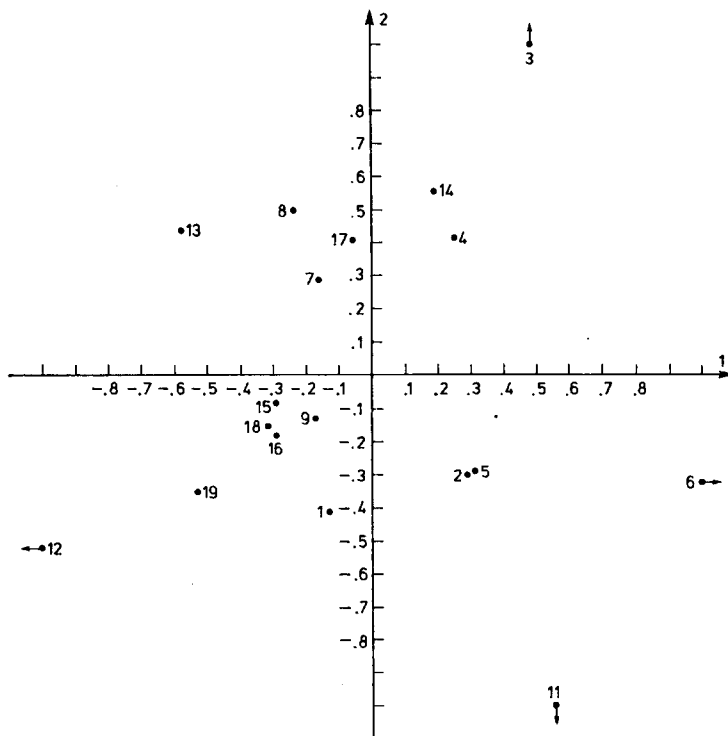
gdzie $t_{i.}$ oraz $t_{.j}$ są liczbami ankietowanych, którzy zaznaczyli czynności A_i oraz A_j . Wskaźniki podobieństwa s'_{ij} są oczywiście symetryczne i przyjmują wartości w przedziale $[0, 1]$, przy czym $s'_{ij} = 0$, gdy $t_{ij} = 0$, oraz $s'_{ij} = 1$, gdy $t_{ij} = t_{i.} = t_{.j}$. W szczególności mamy "podobieństwo do samego siebie": $s'_{ii} = 1$. Za pomocą przekształcenia

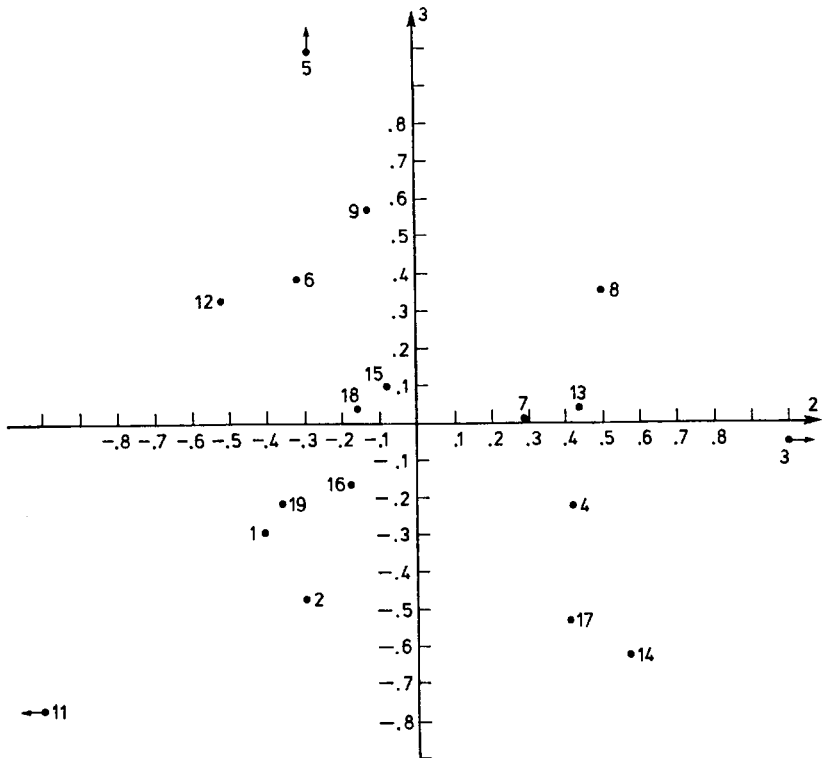
$$(1.7) \quad s_{ij} = 1 - s'_{ij}, \quad i, j = 1, \dots, n,$$

otrzymuje się wskaźniki zróżnicowania s_{ij} . Są one symetryczne, przyjmują wartości z przedziału $[0, 1]$ i spełniają warunek trójkąta (por. [15]).

Symetryczną macierz $S = (s_{ij})$ odległości subiektywnych można przeanalizować metodą skalowania wielowymiarowego i każdej czynności A_i przyporządkować w r -wymiarowej przestrzeni metrycznej punkt X_i tak, żeby odległości $d_{ij} = d(X_i, X_j)$, w odpowiednim metrycznym lub niemetrycznym sensie, korespondowały z wartościami s_{ij} .

Ponieważ "czynność" 20 jest teoretycznie zastępnikiem dla jednej z dziewiętnastu pierwszych czynności i ponieważ czynność 21, ze względu na duże zróżnicowanie zajęć, jakie pod nią można rozumieć, nie może być sensownie kojarzona z żadną z dziewiętnastu pierwszych czynności, obie te czynności wyeliminowano z dalszej analizy; mamy tu więc $n = 19$. Subiektywne odległości $s_{i,10}$ czynności 10 od każdej z pozostałych osiemnastu czynności są większe niż wszelkie wzajemne odległości subiek-





Rys. 1. Wyniki niemetrycznego skalowania wielowymiarowego

tywne pozostałych czynności między sobą: ankietowani, którzy zaznaczali czynność 10 bardzo rzadko wyróżniali jeszcze którąś z pozostałych czynności. W takiej sytuacji metody numeryczne minimalizacji każdego z kryteriów (1.6) prowadzą do rozwiązania "zdegenerowanego", w którym wszystkie czynności, poza czynnością 10, są reprezentowane za pomocą jednego punktu. Z tego powodu z analizy niemetrycznej wyeliminowano czynność 10. Na rysunku 1 przedstawiono wyniki analizy niemetrycznej dla przypadku $r = 3$, w trzech rzutach na płaszczyzny wyznaczone przez odpowiednie osie współrzędnych.

Konfiguracja otrzymana w wyniku tej analizy i przedstawiona na rysunku 1 ma tę własność, że rangi odległości d_{ij} pomiędzy osiemnastu rozważanymi kategoriami różnią się od rang zaobser-

wowanych odległości subiektywnych s_{ij} możliwie mało w sensie unormowanego kryterium (1.5). Konfiguracja ma punkt ciężkości w początku układu współrzędnych, główne jej osie są skierowane zgodnie z osiami układu współrzędnych, a poszczególne współrzędne są unormowane tak, że ich wariancja jest równa jedności. Interpretacja otrzymanych wyników wymaga rozważania wszystkich trzech części rysunku łącznie; bardziej przejrzysty byłby oczywiście niemożliwy tu do pokazania wykres przestrzenny.

Przeprowadźmy najpierw interpretację klasyfikacyjną otrzymanej konfiguracji. Jeżeli w danej konfiguracji można wyróżnić grupy (skupienia, "clusters") punktów bliskich, to można wnioskować, że odpowiednie sposoby spędzania wolnego czasu często współpracują ze sobą. Punkty izolowane reprezentują takie czynności, które rzadko łączą się z innymi czynnościami. Przyglądając się wszystkim trzem częściom rysunku 1, można na początek wyróżnić trzy skupienia punktów:

- skupienie 1 : 15. czytam książki, słucham radia lub muzyki z płyt,
 16. oglądam telewizję,
 1. wykonuję pracę typu rzemieślniczego w domu i w ogródku,
 18. chodzę na spacer, wędrowki, wycieczki,
 19. spędzam wolny czas z dziećmi;
- skupienie 2 : 7. uczęszczam na imprezy rozrywkowe, wieczorki taneczne,
 8. chodzę do kina,
 13. spotykam się z przyjaciółmi;
- skupienie 3 : 4. uczęszczam na imprezy sportowe,
 14. chodzę na piwo,
 17. gram w karty, szachy, itp.

Jako izolowane punkty wyróżniają się:

3. uprawiam sport.

Ta czynność jest najbliższa czynnościom ze skupienia 3 i ewentualnie spokrewniona z czynnościami ze skupienia 2;

5. uczęszczam na wykłady i kursy dokształcające;

6. kolekcjonuję znaczki pocztowe, monety,...

Oba izolowane punkty 5 i 6 są najbardziej bliskie wzajemnie sobie i wykazują pewien związek z zajęciem 9;

11. majsterkuję.

Ta czynność jest najbliższa czynności 2;

12. wykonuję robótki ręczne.

To zajęcie jest dość bliskie zajęciom grupy 1 i być może również grupy 2.

Pozostają dwie czynności:

2. pielęgnuję swój pojazd,

9. chodzę do teatru, na koncerty, wystawy,

których sytuację daje się wyjaśnić tak wyróżnionymi skupieniami jak i wyróżnionymi punktami izolowanymi.

Podobne wyniki można otrzymać za pomocą analizy skupień (por. rozdz. 3), ale dla zorientowanego Czytelnika jest bezpośrednio oczywiste, że metryczna ilustracja dostarcza daleko więcej informacji o analizowanych obiektach niż zwykłe metody analizy skupień.

Przeprowadzona wyżej interpretacja wydaje się na pierwszy rzut oka banalna, ale doświadczenie pokazuje, że twierdzenia, które a posteriori wydają się całkiem "oczywiste", rzadko kiedy mogą być w tej formie podane przez ekspertów a priori. Ponadto rozważany tu przykład został świadomie tak wybrany, żeby wyniki analizy w znacznym stopniu pokrywały się z naszymi wyobrażeniami.

Dalsza interpretacja wyników skalowania wielowymiarowego, a mianowicie tzw. analiza cech (w socjologii używa się terminu "analiza wymiarów") w takiej postaci, w jakiej ją tu przedstawiamy, jest przez różnych badaczy akceptowana tylko z pewnymi zastrzeżeniami. W analizie cech podejmuje się próbę przypisania każdej z trzech zmiennych służących za podstawę układu współrzędnych znaczenia pewnej syntetycznej, ukrytej cechy (skłonności, czynnika) rozważanych obiektów. W naszym przykładzie pierwsza zmienna rozdziela przede wszystkim czynności żeńskie i towarzyskie o pewnym stopniu zażyłości od czynności męskich i indywidualnych; druga zmienna rozdziela przede wszystkim

czynności o charakterze familiarnym i domowym, introwertycznym, od czynności ekstrawertycznych, przekładających towarzystwo innych osób poza swoim własnym domem; w końcu trzecia zmienna rozdziela w naturalny sposób czynności typu robótek ręcznych i w pewnym sensie bezpretensjonalny od czynności duchowo-kulturalnych. To przypisanie poszczególnym współrzędnym znaczenia pewnych cech, które nie są empirycznie bezpośrednio obserwowalne (cech ukrytych!) jest oczywiście nieco kontrowersyjne. Jest interesujące, że filatelistyka (6) jest zajęciem "jak najbardziej męskim", podczas gdy gra w karty (17), chociaż należy do skupienia 3, jest co najmniej tak samo damskim, jak i męskim sposobem spędzania wolnego czasu.

2. WIELOWYMIAROWE SKALOWANIE RÓŻNIC INDYWIDUALNYCH

W praktyce często znajdujemy się w sytuacji, w której chcielibyśmy przeprowadzić skalowanie na podstawie więcej niż jednej macierzy S odległości subiektywnych jednych i tych samych obiektów. Może się na przykład zdarzyć, że dla tych samych obiektów A_1, \dots, A_n otrzymamy od m różnych osób, grup społecznych lub od danej osoby w różnych okresach czasu macierze $S^{(1)} = (s_{ij,1})$, $l = 1, \dots, m$, które będą różniły się między sobą. W takiej sytuacji metody zwykłego skalowania wielowymiarowego mogłyby być użyte na przykład

- a) dla opracowania, niezależnie od siebie, wszystkich m macierzy $S^{(1)}$ i zbudowania m konfiguracji $X^{(1)}$, albo
- b) dla opracowania średniej S^* macierzy zbudowanej z m macierzy indywidualnych $S^{(1)}$ i otrzymania w ten sposób "średniej" konfiguracji X .

Moglibyśmy ewentualnie

- c) zbiór macierzy $S^{(1)}$, $l = 1, \dots, m$, rozbić według pewnego kryterium na jednorodne podzbiory i na podstawie reprezentantów poszczególnych podzbiorów skonstruować odpowiednie konfiguracje.

Oczywiste jest, że jeżeli liczba indywiduów m jest duża i jeżeli postępując jak w a) otrzymamy m konfiguracji, to porów-

nywanie parami tych konfiguracji (ewentualnie odpowiednich wykresów) będzie za mało przejrzyste, aby można było wykryć podobieństwa i różnice pomiędzy indywiduami. Dodatkową trudność przy takich porównaniach sprawia to, że poszczególne konfiguracje są niezmiennicze względem przekształceń, o których już mówiliśmy (w takiej sytuacji odpowiednią analizę może ułatwić "uogólniona metoda prokrustowa" (por. [3], [13]), która polega na tym, aby poszczególne konfiguracje tak przekształcić przez przesunięcia, obroty i jednokładność, aby minimalnie różniły się między sobą). Postępowanie b) ma tę wadę, że zacierają indywidualne zróżnicowanie macierzy $S^{(l)}$, $l = 1, \dots, m$, i pozwala na wydobycie tylko pewnych wspólnych własności.

Ponieważ istnieją liczne sytuacje, w których również postępowanie c) nie jest zadowalające, opracowano specjalne metody jednoczesnego skalowania m macierzy S . Metody te noszą nazwę metod skalowania wielowymiarowego różnic indywidualnych. Ogólnie rzecz biorąc, metody te polegają na wprowadzeniu pewnych dodatkowych parametrów w taki sposób, że zróżnicowanie indywidualne macierzy $S^{(l)}$ opisane zostaje przez różne wartości tych parametrów. Istnieje tu szereg metod różniących się pomiędzy sobą sposobem wprowadzenia tego dodatkowego parametru. Najczęściej stosowana bywa metoda Horana [7], w związku z czym przedstawimy ją tu dokładniej.

W metodzie Horana zakłada się, że konfiguracje $X^{(l)}$, $l = 1, \dots, m$, różnią się, z dokładnością do pewnego błędu losowego, od pewnej wspólnej konfiguracji Y tylko pewnymi mnożnikami w_{lk} :

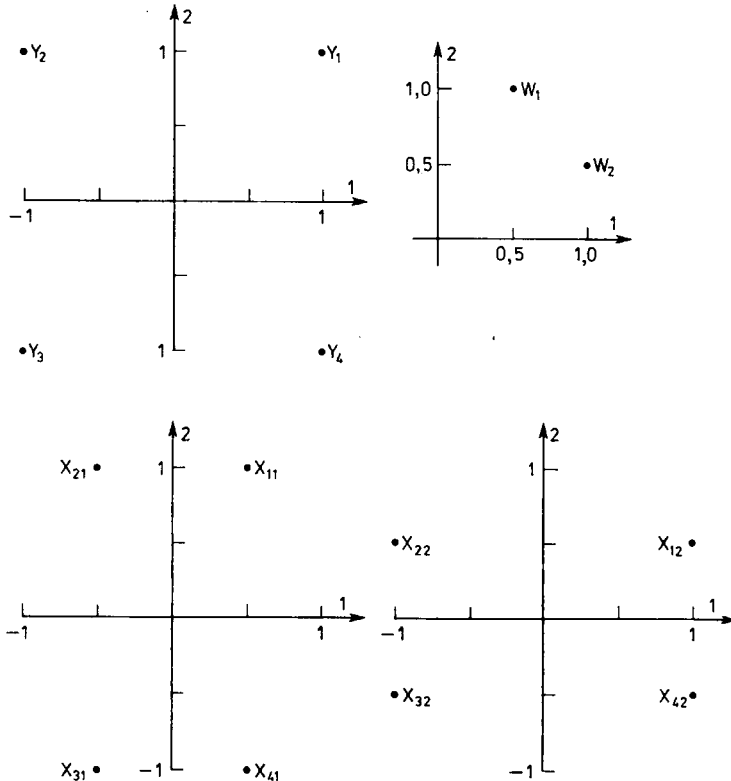
$$(2.1) \quad x_{ik.l} = w_{lk}y_{ik}, \quad i = 1, \dots, n, \quad k = 1, \dots, r, \quad l = 1, \dots, m,$$

gdzie i jest numerem obiektu, k jest numerem współrzędnej oraz l jest numerem konfiguracji: $X^{(l)} = (x_{ik.l})$. Odległość euklidesowa pomiędzy obiektami A_i oraz A_j przyjmuje teraz postać

$$(2.2) \quad d_{ij.1} = \left(\sum_{k=1}^r (x_{ik.1} - x_{jk.1})^2 \right)^{1/2} =$$

$$= \left(\sum_{k=1}^r w_{1k}^2 (y_{ik} - y_{jk})^2 \right)^{1/2}$$

i zależy tylko od współrzędnych y_{ik} punktów wspólnej konfiguracji Y i indywidualnych wag w_{1k} . Znając macierze $Y = (y_{ij})$ oraz $W = (w_{1k})$, można, za pomocą wzoru (2.1), wyznaczyć macierze (konfiguracje) $X^{(1)}$, które w tym kontekście przyjęto nazywać macierzami indywidualnymi. Prostą ilustrację dla $m = 2$ oraz $n = 4$ przedstawia rysunek 2. Jest oczywiste, że w przypadku dużej liczby m badanych macierzy, łatwiej jest wnioskować z dwóch, niż z m konfiguracji (lub odpowiednich rysunków).

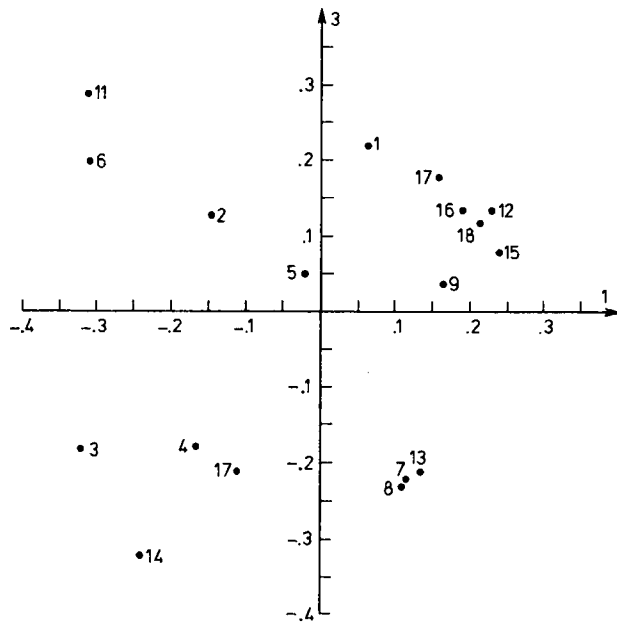
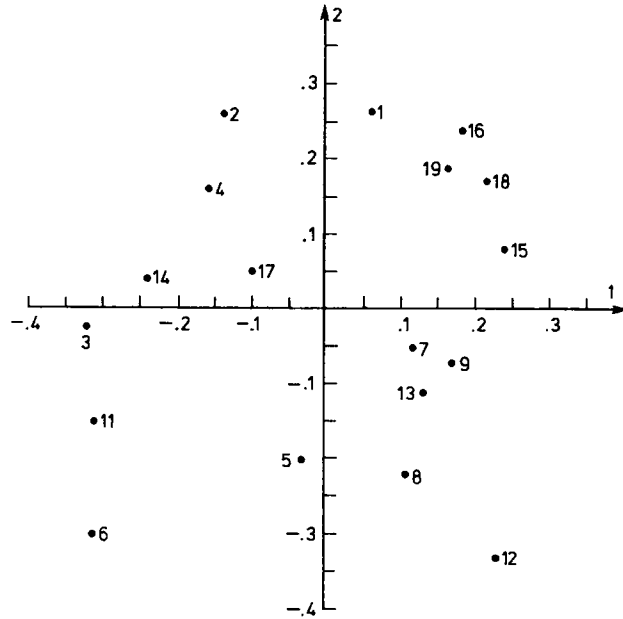


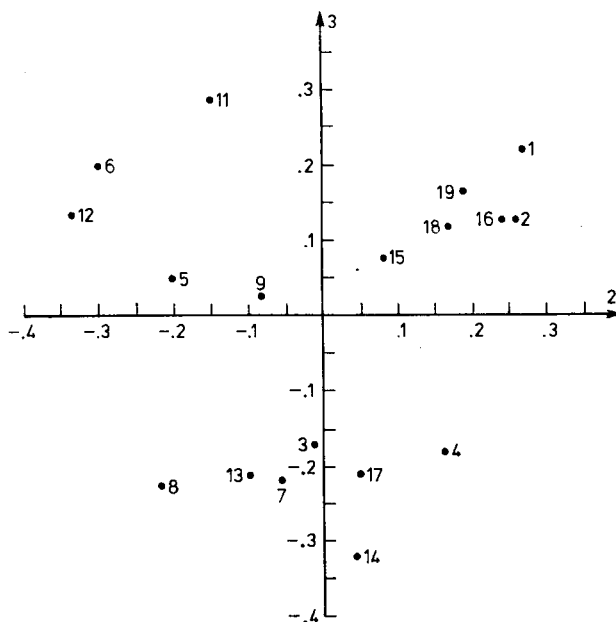
Rys. 2. Związki pomiędzy przestrzeniami Y i W oraz $X^{(1)}$ i $X^{(2)}$ (pierwsze indywiduum rozróżnia obiekty kierując się bardziej niż drugie wartością drugiej cechy, natomiast drugie indywiduum w większym stopniu niż pierwsze uwzględnia wartość pierwszej cechy)

Przy danych macierzach odległości subiektywnych $S^{(l)}$, $l = 1, \dots, m$, i przy danym wymiarze r , zadanie wyznaczenia "optymalnych" macierzy Y oraz W jest analogiczne do rozważanego już zadania w przypadku zwykłej metody skalowania wielowymiarowego. Podobnie wygląda sprawa optymalnego wyboru wymiaru r . Pewien dodatkowy problem stanowi sprawa weryfikacji tego, czy indywidualne zróżnicowanie macierzy S dostatecznie dobrze opisuje się za pomocą wspólnej konfiguracji Y i macierzy wag W . Naturalne pytanie polega tu na tym, czy wydobyto wszystkie istotne informacje o zróżnicowaniu indywidualnym. Jest to znowu pytanie raczej statystycznej niż numerycznej natury. Przy okazji swojego algorytmu COSPA Schönemann [14] podał pewien test, który a posteriori pozwala rozstrzygać, czy macierze $S^{(l)}$, $l = 1, \dots, m$, dają się opisać za pomocą modelu Horana.

Duża popularność modelu Horana wynika prawdopodobnie stąd, że wyniki analizy zaprezentowane w postaci macierzy Y i W są bardzo przejrzyste dla interpretacji. Zakłada się, że subiektywne odległości w macierzach $S^{(l)}$ odzwierciedlają różnice pomiędzy danymi m obiektami w taki sposób, w jaki mogą być one oszacowane przez porównania parami, dokonywane przez m osób, i że osie układu współrzędnych, w którym skonstruowano wspólną konstelację Y , reprezentują określone cechy obiektów. Wtedy dodatnie wagi w_{lk} , $k = 1, \dots, r$, mogłyby być interpretowane jako "stopień znaczenia" k -tej cechy dla l -tej osoby przy ocenie różnicy pomiędzy obiektami: czym większe w_{lk} , tym większą rolę odgrywa w opinii l -tej osoby k -ta cecha przy porównywaniu badanych obiektów.

Interpretacja rozwiązania (Y, W) modelu Horana wymaga, podobnie jak w metodzie skalowania wielowymiarowego, uwzględnienia jego niezmienniczości. Translacja konfiguracji Y nie zmienia odległości $d_{ij,l}$ (por. wzór (2.2)), więc na ogół żąda się takiej normalizacji, aby punkt ciężkości tej konfiguracji leżał w początku układu współrzędnych. W odróżnieniu od zwykłej metody skalowania wielowymiarowego, dowolne ortogonalne rotacje konfiguracji Y prowadzą do zmian odległości $d_{ij,l}$. Zamiast warunku, aby główne osie konfiguracji pokrywały się z osiami układu współrzędnych można tu postulować uporządkowanie kolejności osi układu współrzędnych według malejącej wariancji.

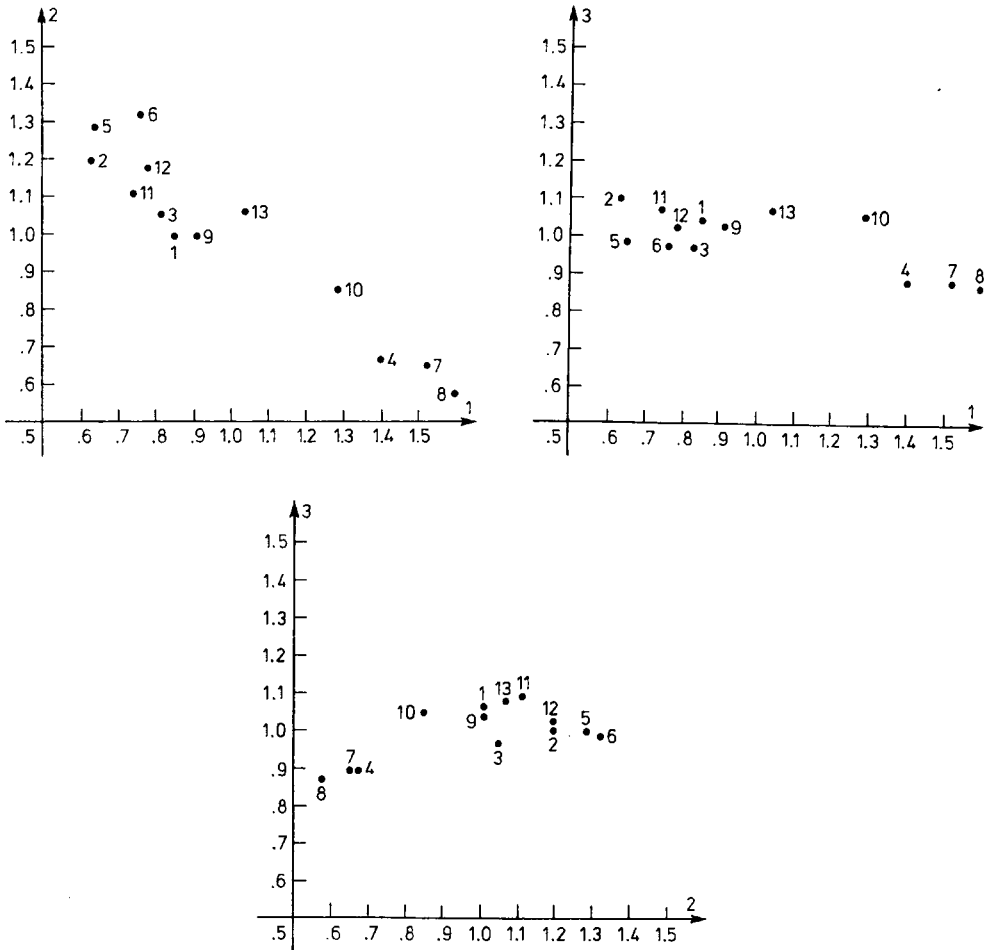




Rys. 3. Macierz konfiguracji Y

W następującym przykładzie prezentujemy tzw. trójdzielcze rozwinięcie przykładu, który dyskutowaliśmy już poprzednio przy okazji metody skalowania wielowymiarowego. Przeanalizujemy metodą Horana macierze $S^{(1)} = (s_{ij.1})$, $i, j = 1, \dots, 18$, dla $m = 13$ grup zawodowych ($l = 1, \dots, 13$), przy czym, jak poprzednio, będą to macierze odległości subiektywnych Tanimoto. Rozważano następujące grupy zawodowe:

1. robotnicy produkcji podstawowej,
2. robotnicy techniczni w procesach pomocniczych,
3. robotnicy inni w procesach pomocniczych,
4. robotnicy w służbie socjalnej i w zaopatrzeniu,
5. brygadziści,
6. mistrzowie i nadmistrzowie,
7. techniczni pracownicy umysłowi bez funkcji kierowniczych,
8. urzędnicy bez funkcji kierowniczych,
9. inżynierjno-techniczni pracownicy umysłowi bez funkcji kierowniczych,



Rys. 4. Macierz wag W

10. ekonomiści bez funkcji kierowniczych,
11. kierownicy pierwszego szczebla,
12. kierownicy drugiego szczebla,
13. kierownicy trzeciego szczebla.

Listę rozważanych sposobów spędzania wolnego czasu podaliśmy już poprzednio (znowu nie uwzględniamy zajęć 10, 20 i 21). Na rysunkach 3 i 4 przedstawiamy (dla $r = 3$) wynikową konfigurację Y oraz wagi W. Analiza macierzy wag W przedstawionej na rysunku 4 sugeruje, że trzynaście grup zawodowych różni się

między sobą prawie wyłącznie przy ocenie pierwszej i drugiej zmiennej układu współrzędnych, w którym otrzymano macierz Y . Wyraźnie widać, że cztery grupy zawodowe: 4, 7, 8 i 10 wyróżniają się spośród innych grup, z których najbliższa jest im grupa 13, a najdalsze są grupy 2, 5 i 6. Konfiguracje $X^{(1)}$ dla grup 1 = 4, 7, 8 i 10 będą różniły się od wspólnej konfiguracji Y tym, że będą w stosunku do niej bardziej rozciągnięte wzdłuż pierwszej i bardziej skupione wzdłuż drugiej osi układu współrzędnych. Podobnie konfiguracje $X^{(2)}$, $X^{(5)}$ oraz $X^{(6)}$ będą bardziej rozciągnięte wzdłuż drugiej osi i bardziej skupione wzdłuż pierwszej.

3. HIERARCHICZNA ANALIZA SKUPIEŃ

Jak już wspominaliśmy w pierwszym rozdziale, część wyników, które otrzymaliśmy metodą skalowania wielowymiarowego, można otrzymać lub potwierdzić metodami analizy skupień. Objaśnimy krótko jedną z metod hierarchicznej analizy skupień, a następnie pokażemy, że klasyfikacja otrzymana w pierwszym rozdziale jako wniosek z analizy konfiguracji skonstruowanej niemetryczną metodą skalowania wielowymiarowego w zasadzie pokrywa się z klasyfikacją otrzymaną za pomocą hierarchicznej analizy skupień (dla tych samych danych S).

Celem analizy skupień jest podzielenie skończonej liczby n badanych obiektów A_1, \dots, A_n na pewną liczbę m klas (skupień) C_1, \dots, C_m w taki sposób, żeby każde dwa obiekty z jednej klasy były bardzo do siebie podobne, a każde dwa obiekty z różnych klas różniły się znacznie między sobą. Znane jest dużo różnych metod analizy skupień różniących się między sobą przede wszystkim sposobem mierzenia jednorodności wewnątrzklasowej i zróżnicowania między klasami na podstawie danych o obiektach. Wiele z tych metod za punkt wyjścia przyjmuje symetryczną macierz $S = (s_{ij})$ stopnia $n \times n$, w której s_{ij} jest wielkością wyrażającą stopień zróżnicowania lub odległość pomiędzy obiektami A_i i A_j , a więc takie same dane, jakie rozważaliśmy w pierwszym rozdziale przy okazji zwykłej metody skalowania wielowymiarowego. Cho-

ciaż metoda skalowania wielowymiarowego i metoda analizy skupień operują tą samą macierzą S , ich cele i metody rachowania są bardzo odmienne.

Aglomerująca hierarchiczna analiza skupień prowadzi do wyznaczenia n różnych podziałów zbioru n obiektów na skupienia w taki sposób, że każdy z n obiektów zostaje zaliczony do dokładnie jednego z m skupień danego podziału. Te podziały opisuje się za pomocą pewnej hierarchii: na dole i na górze hierarchii znajdują się pewne szczególne podziały, a przy każdym przejściu z jednego poziomu hierarchii na poziom wyższy dwa skupienia zostają połączone w jedno. Na najniższym poziomie hierarchii znajduje się podział na n skupień (z których każde składa się dokładnie z jednego obiektu), a na najwyższym - tylko jedno skupienie zawierające wszystkie n obiektów. Przypuśćmy, że na pewnym poziomie hierarchii zbiór n obiektów jest podzielony na m skupień C_1, \dots, C_m i że dana jest macierz $S = (s_{pq})$, $p, q = 1, \dots, m$, odległości s_{pq} pomiędzy skupieniami C_p i C_q . Na najniższym poziomie hierarchii macierz ta pokrywa się z macierzą danych $S = (s_{ij})$. Dla wyznaczenia nowych skupień i obliczenia odległości s_{pq} na kolejnych poziomach hierarchii proponowano liczne sposoby; przedstawimy jeden z częściej używanych. Przy przejściu na wyższy poziom hierarchii łączy się ze sobą takie skupienia C_u i C_v , dla których s_{uv} jest najmniejsze, pozostawiając wszystkie pozostałe skupienia bez zmiany. Niech $C_w = C_u \cup C_v$ będzie nowopowstałym skupieniem. Dla obliczenia odległości s_{wp} , $p \neq u$, $p \neq v$, nowego skupienia od pozostałych Späth [15] podaje aż siedem różnych wzorów, na przykład

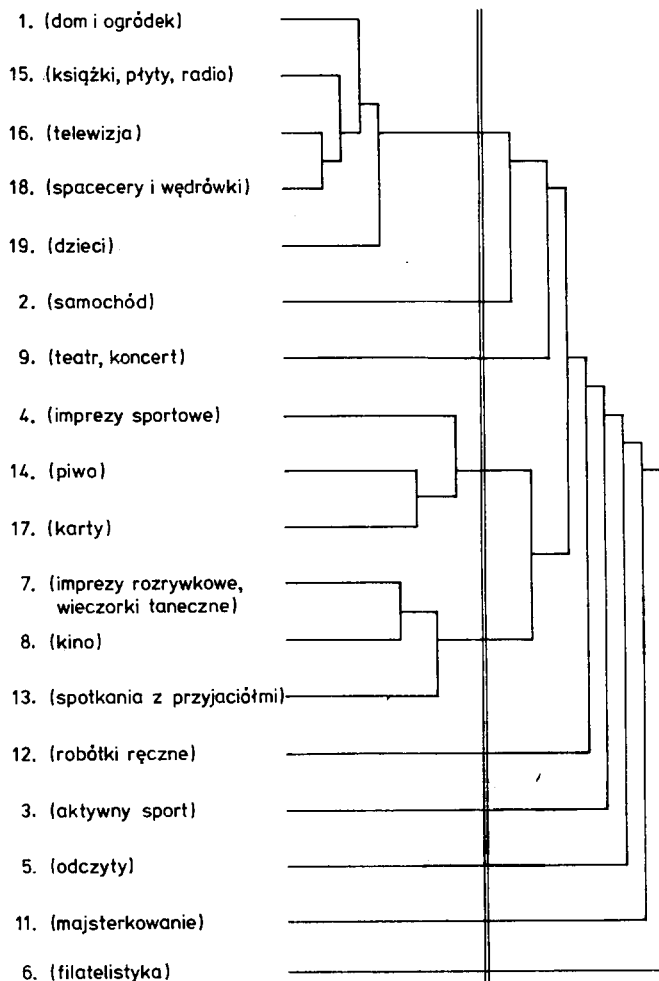
$$(3.1) \quad s_{wp} = \min \{s_{up}, s_{vp}\},$$

$$(3.2) \quad s_{wp} = \max \{s_{up}, s_{vp}\},$$

$$(3.3) \quad s_{wp} = (n_u s_{up} + n_v s_{vp}) / n_w,$$

$$(3.4) \quad s_{wp} = (s_{up} + s_{vp}) / 2,$$

gdzie n_p jest liczbą punktów w skupieniu C_p .



Rys. 5. Dendryt hierarchicznej analizy skupień

Dla przykładu przedstawimy wyniki analizy danych zawartych w macierzy S z rozdziału 1 za pomocą hierarchicznej analizy skupień z algorytmem (3.3). Po ośmiu kolejnych krokach aglomeracji otrzymano wyniki przedstawione w postaci dendrytu na rysunku 5. Wyróżniły się tu skupienia:

1, 15, 16, 18, 19
2
9
4, 14, 17
7, 8, 13
12
3
5
11
6

co dokładnie pokrywa się z wynikami analizy metodą skalowania wielowymiarowego opierającej się na konfiguracji przedstawionej na rysunku 1. Ponieważ metody skalowania wielowymiarowego i analizy skupień są w znacznym stopniu niezależne od siebie, taka zgodność rezultatów sugeruje, że obie metody w podobny sposób opisują informacje zawarte w rozważanej macierzy danych S .

PRACE CYTOWANE

- [1] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, H.D. Brunk, Statistical Inference under Order Restrictions, J. Wiley & Sons, New York 1972.
- [2] J.D. Carroll, J.J. Chang, Analysis of individual differences in multidimensional scaling via an N-way generalization of ECKARTYOUNG decomposition, Psychometrika 35 (1970), 283-319.
- [3] J.C. Gower, Generalized Procrustes Analysis, Psychometrika 40 (1975), 33-51.

- [4] J.A. H a r t i g a n, Clustering Algorithms, J. Wiley & Sons, New York 1975.
- [5] W. H a r t m a n n, Geometrische Modelle zur Analyse empirischer Daten, Akademie Verlag, Berlin 1979.
- [6] W. H a r t m a n n, R. W e l s k o p f, Zu Anwendungen der multidimensionalen Skalierung in der Soziologie, Jahrbuch für Soziologie und Sozialpolitik 1 (1980), 181-192.
- [7] C.B. H o r a n, Multidimensional scaling: Combining observations when individuals have different perceptual structures, Psychometrika 34 (1969), 139-165.
- [8] J.B. K r u s k a l, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964), 1-27.
- [9] J.B. K r u s k a l, Nonmetric multidimensional scaling: A numerical method, Psychometrika 29 (1964), 115-129.
- [10] J.C. L i n g o e s, Some boundary conditions for a monotone analysis of symmetric matrices, Psychometrika 36 (1971), 195-203.
- [11] J.O. R a m s a y, Maximum Likelihood Estimation in Multidimensional Scaling, Psychometrika 42 (1977), 241-276.
- [12] P.H. S c h ö n e m a n n, An algebraic solution for a class of subjective metrics models, Psychometrika 37 (1972), 441-451.
- [13] P.H. S c h ö n e m a n n, R.M. C a r r o l l, Fitting one matrix under choice of a central dilation and a rigid motion, Psychometrika 35 (1970), 245-255.
- [14] P.H. S c h ö n e m a n n, F.S. C a r t e r, W.L. J a m e s (1976), Contributions to subjective metrics scaling:
- I. COSPA, a fast method for fitting and testing HORAN's model, and an empirical comparison with INDSICAL and ALSCAL;
- II. A statistical test and approximate norms and evaluating the fit of HORAN's model with COSPA;
- Unveröffentlichte Papiere, PURDUE University.

-
- [15] H. S p ä t h, Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion, Oldenbourg, München 1975.
- [16] Y. T a k a n e, F.W. Y o u n g, J. d e L e e u w, Non-metric individual differences scaling: An alternating least squares method with optimal scaling features, Psychometrika 42 (1977), 7-67.
- [17] W.S. T o r g e r s o n, Theory and Methods of Scaling, J. Wiley & Sons New York 1958.