POLSKA AKADEMIA NAUK
INSTYTUT GEOGRAFII
I PRZESTRZENNEGO ZAGOSPODAROWANIA

ZESPÓŁ KOORDYNACYJNY PROBLEMU MIĘDZYRESORTOWEGO
„PODSTAWY PRZESTRZENNEGO ZAGOSPODAROWANIA KRAJU"

Do użytku służbowego

BIULETYN INFORMACYJNY

ZESZYT 17

FRANCISZEK A. SZCZOTKA

PODSTAWY TAKSONOMII NUMERYCZNEJ

WARSZAWA 1976

http://rcin.org.pl

POLSKA AKADEMIA NAUK
INSTYTUT GEOGRAFII
I PRZESTRZENNEGO ZAGOSPODAROWANIA

ZESPÓŁ KOORDYNACYJNY PROBLEMU MIĘDZYRESORTOWEGO
„PODSTAWY PRZESTRZENNEGO ZAGOSPODAROWANIA KRAJU"

Do użytku służbowego

BIULETYN INFORMACYJNY

ZESZYT 17

FRANCISZEK A. SZCZOTKA

PODSTAWY TAKSONOMII NUMERYCZNEJ

- wykład -

WARSZAWA 1976

http://rcin.org.pl

Opracowanie redakcyjne: Irena Stańczak

W.D.N. Zam. 692/o/76. Nakład 200+23 egz.

http://rcin.org.pl

SPIS TREŚCI

strona

Przedmowa - K. Bielecka 5

1. Podstawowe zagadnienia taksonomii 9

1.1 Zadania taksonomii 9

1.2 Podstawowe ustalenia taksonomii 11

1.2.1 Miary podobieństwa 11

1.2.2 Kryterium dobroci podziału 13

1.2.3 Wybór algorytmu podziału 14

1.3 Podziały 14

1.4 Przykład procedury porządkowania i podziału: Taksonomia Wrocławska 16

2. Miary podobieństwa 19

2.1 Uwagi wstępne 19

2.2 Właściwości wskaźników podobieństwa 20

2.3 Przykłady wskaźników podobieństwa 21

2.3.1 Wskaźniki podobieństwa w przypadku cech alternatywnych 21

2.3.2 Wskaźniki podobieństwa w przypadku cech wielodzielczych 28

2.4 Miary podobieństwa dla cech wielkościowych 29

2.5 Związki formalne między wskaźnikami podobieństwa a odległościami 37

http://rcin.org.pl

strona

3. Kryteria optymalności podziałów 38

3.1 Wstęp 38

3.2 Miary jednorodności wewnątrz grup 39

3.3 Miary podobieństwa grup 41

3.4 Kryteria dobroci podziału 42

4. Procedury grupowania 44

4.1 Wstęp 44

4.2 Przykłady procedur grupowania 45

4.2.1 Wprowadzenie 45

4.2.2 Metoda diagramu Czekanowskiego 46

4.2.3 Metoda najbliższego sąsiada 48

4.2.4 Metoda najdalszego sąsiada 49

4.2.5 Metody podziałowe Huberta 51

4.2.6 Metoda łączonych par Sokala i Michenera 57

4.3 Grupowanie przy z góry zadanym kryterium optymalności 60

5. Porządkowanie liniowe 61

6. Redukcja ilości zmiennych 63

Literatura cytowana 67

Spis tablic 69

http://rcin.org.pl

PRZEDMOWA

Podejmowanie coraz bardziej skomplikowanych problemów badawczych stwarza potrzebę uściślenia wyników badań, a co za tym idzie, konieczność stosowania w coraz większym zakresie metod ilościowych. Potrzebom tym wychodzi naprzeciw bujnie rozwijająca się matematyka, dysponująca arsenałem licznych nowych metod i technik.

Wykorzystanie metod matematycznych jako nowego narzędzia badawczego w danej dziedzinie jest na ogół przedsięwzięciem trudnym i kryjącym w sobie niebezpieczeństwo popełniania nieprawidłowego "umatematyczniania" badań, co w konsekwencji prowadzi do przyjmowania i interpretacji merytorycznie fałszywych wyników jako dobrych - czego wiarę pokłada się w tym, że zostały one uzyskane przy użyciu obiektywnych i precyzyjnych metod.

Wobec realności takiego zagrożenia, szczególnie nowym próbom zastosowania metod ilościowych winna towarzyszyć świadomość, że uzyskanie poznawczo pozytywnych wyników uwarunkowane jest wnikliwym potraktowaniem merytorycznej strony badanego zagadnienia w powiązaniu z rozpatrzeniem możliwości adaptacyjnych aparatu metod ilościowych. Prawidłowe wyznaczenie jakościowych kryteriów badanego zagadnienia oraz właściwy dobór materiału obserwacyjnego, na którym mają być zastosowane metody ilościowe ma dla strony poznaw-

http://rcin.org.pl

czej zasadnicze znaczenie, żadna bowiem nawet najbardziej precyzyjna metoda ilościowa nie koryguje popełnionych w tym zakresie błędów (niedopracowań). Myśl tę uplastycznia przytaczane przez znakomitego uczonego A.N. Kryłowa porównanie matematyki z kamieniem młyńskim: "miele on to co się pod niego wsypie, a jeśli wsypie się lebiodę, to mąki nie będzie".

Przy prawidłowości jakościowej materiału obserwacyjnego efektywność zastosowania metod ilościowych zależy od trafności wyboru metody najlepszej, spośród właściwych, dla rozwiązania danego zagadnienia. Wymaga to od badacza, by dysponował on odpowiednim zasobem wiedzy o skuteczności działania określonych metod.

W przekonaniu o celowości szerszego stosowania metod ilościowych w badaniach przestrzennych i mając zarazem na uwadze związane z tym trudności oraz wspomniane grożące niebezpieczeństwa, podjęte zostały w ramach prac Grupy Tematycznej 03: Struktura przestrzenna wyżywienia i rolnictwa, Problemu węzłowego 11.2.1.: Podstawy przestrzennego zagospodarowania, prace nad uprzystępnieniem badaczom zajmującym się problematyką przestrzenną - wiedzy o wybranych, potencjalnie właściwych dla tej dziedziny, metodach ilościowych.

Do metod takich niewątpliwie należą m.in. metody taksonomiczne służące dokonywaniu klasyfikacji i porządkowaniu empirycznego meteriału obserwacyjnego.

Istotnym z punktu widzenia badań przestrzennych walorem tych metod jest to, że pozwalają one klasyfikować (dzielić na grupy, określać podobieństwo, czy stwierdzać niepodobieństwo) obiekty charakteryzowane za pomocą więcej niż jednej cechy. Właściwość ta ma doniosłe znaczenie dla zakresu poznania, gdyż daje ona możność uwzględniania w opisie badanego obiektu (zjawiska) wszystkich, bądź większości, charakterystycznych dla niego cech i dokonywania

http://rcin.org.pl

analiz podobieństwa obiektów - ze względu na podobieństwo struktur rozpatrywanych cech. W związku z tym metody taksonomiczne mogą mieć zastosowanie we wszelkich systematykach i typologiach, które, jak wiadomo, z istoty swej opierają się zazwyczaj na licznych charakterystykach (cechach). Wykorzystanie metody typologii jako narzędzia analizy w badaniach przestrzennych stwarza tym badaniom szanse pełniejszego poznania złożonej rzeczywistości.

Opracowanie niniejsze stanowi wykład podstaw taksonomii numerycznej. Omówione w nim zostały: podstawowe zagadnienia taksonomii, miary podobieństwa, kryteria optymalności podziału, procedury grupowania i porządkowania. Napisany językiem przystępnym, ilustrowany łatwymi przykładami zastosowań, adresowany przede wszystkim do badaczy mniej zaawansowanych w znajomości matematyki - oby służył z pożytkiem.

Krystyna Bielecka

Warszawa, 20 lutego 1976 r.

http://rcin.org.pl

http://rcin.org.pl

1. PODSTAWOWE ZAGADNIENIA TAKSONOMII

1.1 Zadania taksonomii

Przez taksonomię rozumie się ogólnie naukę o porządkowaniu i klasyfikowaniu. Przedmiotem taksonomii są podstawy teoretyczne porządkowania i klasyfikacji, zasady ogólne na jakich klasyfikacja i porządkowanie mają być przeprowadzone oraz szczegółowe metody i reguły postępowania.

Jednostki klasyfikowane mogą być różnego rodzaju. Mogą to być osoby, przedmioty, zbiory osób (np. rodziny), zbiory przedmiotów, jednostki natury organizacyjnej (np. województwa), zdarzenia (np. choroby), zjawiska (np. klimatyczne) lub też cechy (np. wielkość produkcji określonego towaru).

Jednostki klasyfikowane nazywać będziemy obiektami, nawet jeżeli są to zdarzenia lub cechy. Oznaczać je będziemy symbolami P_1, P_2, ..., P_n, gdzie n jest ich liczbą. Przez taksonomię numeryczną będziemy rozumieli naukę o porządkowaniu i klasyfikacji opartą na takich informacjach ilościowych o klasyfikowanych obiektach, które pozwalają stosować metody numeryczne.

Przez porządkowanie będziemy rozumieli sformalizowane przedstawienie struktury podobieństwa między obiektami badanego zbioru. Przez klasyfikację natomiast podział niejednorodnego zbioru obiektów na grupy jednorodne obiektów podobnych ze względu na badane

http://rcin.org.pl

cechy. Podziału należy dokonać tylko na podstawie informacji o zbiorze badanych obiektów, a więc bez korzystania z wcześniejszych wzorów.

Rozważa się dwa typy podziałów: podział na grupy rozłączone i podział na grupy przecinające się, tzn. taki podział, przy którym niektóre obiekty mogą należeć równocześnie do dwu lub więcej grup. W naszym wykładzie ograniczymy się do podziałów rozłączonych.

Potrzeba klasyfikacji i porządkowania pojawia się już na samym początku rozwoju nowej dziedziny badań naukowych. Klasyfikacja ma wprowadzić ład i porządek w gromadzącym się empirycznym materiale obserwacyjnym. Przeprowadza się ją przede wszystkim po to, by odzwierciedlić strukturę podobieństwa między badanymi obiektami, co z kolei pozwala na formułowanie hipotez dotyczących przyczyn formalnych i merytorycznych podobieństw wewnątrz grup i niepodobieństw między grupami.

Wyodrębnione w procesie klasyfikacji grupy jednorodne mogą być w wielu przypadkach uważane za typy.

Metody matematyczne, które służą do rozwiązania zadania klasyfikacji noszą różne nazwy: analiza grup, analiza klastrów (cluster analysis), analiza skupień, klasyfikacja automatyczna itp.

Taksonomia numeryczna jest obecnie dziedziną bardzo szybko rozwijającą się. Jest to spowodowane zarówno coraz większym zastosowaniem ilościowych metod badawczych w naukach nietechnicznych jak i szybkim rozwojem elektronicznej techniki obliczeniowej. Liczba publikacji poświęcona taksonomii numerycznej idzie w setki. Z pozycji książkowych wymienić należy wydaną przez PWN książkę Z. Chojnickiego i T. Czyż (1973) "Metody taksonomii numerycznej w regionalizacji geograficznej". Opracowania o charakterze mono-

http://rcin.org.pl

grafii z zakresu taksonomii numerycznej i klasyfikacji napisali Anderberg (1958), Baumann (1971), Bock (1974), Cole (1969). W opracowaniach tych znaleźć można bogatą bibliografię z tego zakresu.

1.2 Podstawowe ustalenia taksonomii

Jak wspomniano wcześniej przez klasyfikację będziemy rozumieli podział niejednorodnego zbioru obiektów na klasy lub grupy obiektów podobnych. Rozwiązanie tego zagadnienia wymaga pewnych podstawowych ustaleń, które dotyczą:

- wyboru miary podobieństwa między badanymi obiektami,
- wyboru kryterium dobroci podziału,
- wyboru algorytmu podziału.

Ponadto potrzebna jest zasada ustalenia ilości klas lub grup.

1.2.1 Miary podobieństwa

Przez miarę podobieństwa między dwoma obiektami A_k i A_l rozumie się funkcję $\mu (A_k, A_l)$, która każdej takiej parze przyporządkowuje liczbę. W zależności od sposobu określenia funkcji μ otrzymuje się różne miary podobieństwa.

Miary podobieństwa w większości przypadków obliczone są na podstawie cech (jakościowych lub wielkościowych) charakteryzujących badane obiekty, chociaż czasem stopień podobieństwa między każdą parą obiektów oceniany jest bezpośrednio przez badającego.

Miary podobieństwa można podzielić na dwie grupy. Do pierwszej należą tzw. wskaźniki podobieństwa, które oznaczone będą ogólnie przez $s(A_k, A_l)$. Wskaźniki podobieństwa charakteryzują się tym, że większe ich wartości świadczą o większym podobieństwie. Przykładem jest wskaźnik systematyczny używany do wyrażania podobieństwa między zespołami roślin:

http://rcin.org.pl

$$s(A_k, A_l) = \frac{1}{2}\left(\frac{m}{n_k} + \frac{m}{n_l}\right),$$

gdzie m jest liczbą gatunków wspólnych w jednym i drugim zespole, natomiast n_k i n_l są odpowiednio równe ilościom gatunków w zespole A_k i A_l. Wskaźnik ten przyjmuje wartość 0, gdy oba zespoły nie mają wspólnych gatunków (kompletne niepodobieństwo) oraz wartość 1, gdy w obu zespołach rosną te same gatunki (kompletne podobieństwo). Innym wskaźnikiem tego samego podobieństwa jest wskaźnik

$$s(A_k, A_l) = \frac{m}{n_k + n_l - m}$$

Wskaźników podobieństwa używa się zwykle wtedy, gdy obiekty scharakteryzowane są występowaniem lub niewystępowaniem pewnych właściwości.

Drugą grupę miar podobieństwa stanowią abstrakcyjnie rozumiane miary odległości między obiektami. Posługujemy się tutaj następującymi analogiami geometrycznymi. Obiekt scharakteryzowany wartościami p cech $x_1, x_2, \ldots, x_p$ interpretuje się jako punkt A o współrzędnych równych wartościom tych cech w przestrzeni p-wymiarowej. Przez odległość rozumie się funkcje, przyporządkowujące każdej parze takich punktów liczbę nieujemną. Odległość między obiektami A_k i A_l oznaczać będziemy przez $d(A_k, A_l)$. Im obiekty A_k i A_l są podobniejsze, tym odległość między nimi jest mniejsza. Zupełnemu podobieństwu odpowiada odległość zero. Odległości używa się zwykle wtedy, gdy obiekty scharakteryzowane są cechami wielkościowymi.

Jeśli sklasyfikowanym zbiorem obiektów jest zespół cech dodatnio skorelowanych, to miarą podobieństwa między parą cech może być współczynnik korelacji. Jest to wskaźnik podobieństwa.

http://rcin.org.pl

Wskaźniki podobieństwa lub odległości dla wszystkich par zestawia się w tablicy:

	A_1	A_2	A_3	...	A_n
A_1	s_{11}	s_{12}	s_{13}	...	s_{1n}
A_2	s_{21}	s_{22}	s_{23}	...	s_{2n}
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
A_n	s_{n1}	s_{n2}	s_{n3}	...	s_{nn}

gdzie n oznacza ilość obiektów.

Taka tablica wskaźników podobieństw lub odległości jest podstawowym materiałem liczbowym do procedur grupowania.

Zagadnienie ustalania miary podobieństw polega na wyborze wzoru określającego podobieństwo lub odległości.

1.2.2 Kryterium dobroci podziału

Ponieważ celem klasyfikacji jest podział zbioru obiektów na grupy obiektów podobnych, ustalony powinien być sposób mierzenia jednorodności grup. Miara ta powinna równocześnie służyć jako kryterium rozstrzygające, który z dwu podziałów tego samego zbioru obiektów na tę samą ilość grup jest lepszy.

Przy podziałach na grupy nierozłączne używa się kryteriów uwzględniających tylko jednorodność wewnątrz grup. Przy podziale na grupy rozłączne o jakości podziału świadczy dodatkowo odległość (lub niepodobieństwo) wzajemna grup od siebie.

Nie zawsze kryterium jakości podziału jest formułowane w sposób jawny i formalny. Bardzo często tkwi ono w samej procedurze podziału. Dlatego też na ogół porównywanie wyników, jakie dają

http://rcin.org.pl

różne procedury podziału, jest w istocie porównywaniem różnych kryteriów dobroci podziału na podstawie wyników uzyskanych przy ich stosowaniu.

1.2.3 Wybór algorytmu podziału

Znanych algorytmów grupowania jest dzisiaj setki. Można je podzielić przede wszystkim na trzy grupy. Pierwsza grupa, to algorytmy opisane formalnie, w których kryterium klasyfikacji nie jest jawnie sprecyzowane. Te algorytmy są zwykle najprostsze w zastosowaniach.

Druga grupa, to algorytmy stosowane dla jawnie sformułowanych kryteriów optymalności. Nie wszystkie z nich gwarantują uzyskanie rzeczywiście podziału optymalnego, ale sądzić można, że uzyskane tymi metodami rozwiązania nie odbiegają wiele od optymalnych.

Trzecia w końcu kategoria, to procedury szukające optymalnych rozwiązań dla dowolnie sformułowanych kryteriów. Metody te zwykle działają na zasadzie kolejnych przybliżeń.

1.3 Podziały

Przypuśćmy, że chcemy podzielić na grupy obiekty $A_1, A_2, \ldots, A_n$. Przez $G_i^{(m)}$ oznaczamy i-tę grupę przy podziale na m grup. Obiekty należące do tej grupy oznaczamy ogólnie przez $A_{i1}, A_{i2}, \ldots, A_{in_i}$. Symbol n_i oznacza liczbę obiektów w grupie $G_i^{(m)}$. Dowolny podział obiektów na m grup oznaczamy przez $P^{(m)}$ lub $\left\{ G_1^{(m)}, G_2^{(m)}, \ldots, G_m^{(m)} \right\}$.

Fakt, że dowolny obiekt A należy do grupy $G_i^{(m)}$, zapisujemy przez $A \in G_i^{(m)}$. Połączenie dwóch grup $G_s^{(m)}$ i $G_t^{(m)}$, gdzie $s < t$, oznaczamy przez $G_s^{(m)} \cup G_t^{(m)}$. Powstała nowa grupa oznaczo-

http://rcin.org.pl

na będzie przez $G_s^{(m-1)}$. Górny wskaźnik zmniejszył się o jeden z m na m-1, bo przez połączenie dwóch grup ich ilość zmniejszyła się o jeden.

Niech $G_u^{(1)}$ będzie dowolną grupą przy podziale na 1 grup, a $G_v^{(m)}$ dowolną grupą przy podziale na m 1 grup. Grupę złożoną z obiektów należących równocześnie i do $G_u^{(1)}$ i do $G_v^{(m)}$ oznaczać będziemy przez $G_u^{(1)} \cap G_v^{(m)}$.

Zbiór wszystkich obiektów rozpatrywany jako grupa zapisujemy w postaci $G_1^{(1)} = (A_1, A_2, \ldots, A_n)$. Stanowi ona jednocześnie podział na jedną grupę $P^{(1)} = \left\{ G_1^{(1)} \right\}$. Podział na n grup ma postać: $P^{(n)} = \left\{ G_1^{(n)}, G_2^{(n)}, \ldots, G_n^{(n)} \right\} = \left\{ (A_1), (A_2), \ldots, (A_n) \right\}$.

Można rozpatrywać kolejno podziały $P^{(1)}, P^{(2)}, P^{(3)}, \ldots, P^{(n)}$ na jedną, dwie i tak dalej aż do n grup. Będziemy wtedy mieli ciąg podziałów.

Wśród ciągów podziałów ważną rolę odgrywają te, w których każdy następny podział powstaje przez podzielenie jednej z grup. Taki ciąg podziałów nazywa się hierarchią podziałów. Hierarchia podziałów ma także tę właściwość, że zmniejszanie ilości grup odbywa się tylko przez połączenie dwóch grup.

Hierarchia podziałów jest to więc ciąg podziałów $P^{(1)}, P^{(2)}, \ldots, P^{(n)}$ spełniający warunek, że dla każdego $u < v$ oraz $i \leq j$ $G_i^{(u)} \cup G_j^{(v)} = G_i^{(u)}$ lub $G_i^{(u)} \cap G_j^{(v)} = 0$

Hierarchia podziału jest pewnym uporządkowaniem badanego zbioru obiektów. Można ją przedstawić graficznie za pomocą tzw. dendrogramu.

http://rcin.org.pl

1.4 Przykład procedury porządkowania i podziału: Taksonomia Wrocławska

W niniejszej części przedstawiona zostanie prosta, a przy tym efektywna metoda podziału, tak zwana taksonomia wrocławska (zob. Perkal, 1953). Danymi wyjściowymi do tej metody porządkowania jest tablica odległości (lub tablica wskaźników podobieństwa).

Procedura składa się z dwóch etapów:
pierwszy etap polega na konstrukcji najkrótszego dendrytu,
drugi etap na kolejnym podziale tego dendrytu na części.

Dendryt jest to połączenie wszystkich obiektów, spełniające następujące warunki:

a) każdy obiekt jest połączony z co najmniej jednym innym obiektem,
b) można po połączeniach przejść od dowolnego obiektu do każdego innego (spójność dendrytu),
c) każda para obiektów połączona jest dokładnie jedną drogą (dendryt nie zawiera łamanych zamkniętych).

Dendryt wrocławski jest dendrytem najkrótszym, tzn. łączy się obiekty w ten sposób, by suma długości połączeń była możliwie mała. Dzieląc dendryt według określonych kryteriów na części otrzymuje się podział obiektów na grupy.

Konstrukcja najkrótszego dendrytu przebiega następująco. W tablicy odległości w kolumnie odpowiadającej pierwszemu obiektowi szukamy obiektu najbardziej mu podobnego, tzn. najbliższego. Niech będzie nim obiekt A_{j_1} o numerze j_1. Na szkicu roboczym zaznaczamy ten fakt przez narysowanie kreski łączącej obiekt A_1 z obiektem A_{j_1}. Przechodzimy do kolumny odpowiadającej obiektowi A_{j_1} i szukamy w niej obiektu najbardziej mu podobnego. Mogą zajść

http://rcin.org.pl

dwie możliwości. Pierwsza, że najbliższym obiektem będzie obiekt A_{j_2} różny od obiektu pierwszego. Na szkicu roboczym łączymy wtedy kreską obiekt A_{j_2} z A_{j_1} i przechodzimy do kolumny odpowiadającej obiektowi A_{j_2}. Drugi przypadek, jaki się może zdarzyć, to ten, że najbliższy do obiektu A_{j_1} jest obiekt A_1. Poprzestajemy wtedy na tych dwóch obiektach i zaczynamy postępowanie od pierwszego z kolei obiektu dotychczas nierozpatrywanego.

Ogólne postępowanie jest następujące. Niech A_k będzie ostatnim dołączonym obiektem. W kolumnie odpowiadającej temu obiektowi szukamy obiektu najbardziej mu podobnego. Jeżeli obiektem tym jest obiekt poprzednio już połączony przechodzimy do następnego nierozpatrywanego dotychczas obiektu. Jeżeli natomiast obiekt nie był dotąd jeszcze rozpatrywany łączy się go na szkicu z poprzednim i przechodzi do odpowiadającej mu kolumny tablicy odległości. Po rozpatrzeniu wszystkich obiektów otrzymuje się zwykle kilka grup połączonych ze sobą obiektów (rys. 1). Grupy te nazywać będziemy połączeniami rzędu pierwszego. Grupy te łączy się przez parę obiektów najbliższych, z których jeden należy do jednej grupy, drugi do drugiej.

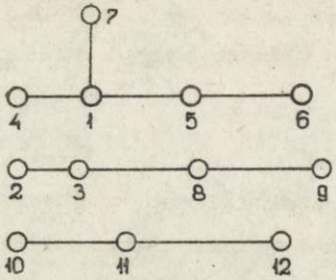
Szkic roboczy przerysowuje się następnie w postaci przejrzystej tak by odcinki łączące obiekty były proporcjonalne do odległości między łączonymi obiektami (rys. 2).

Podział dendrytu na części odbywa się przez odrzucenie kolejne najdłuższych w nim odcinków (rys. 3).

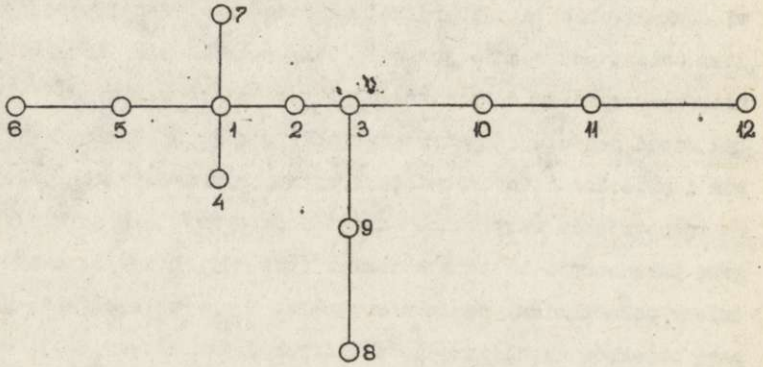
Dla zilustrowania tej metody wzięto 12 powiatów opisanych 21 cechami charakteryzującymi produkcję rolną:

http://rcin.org.pl

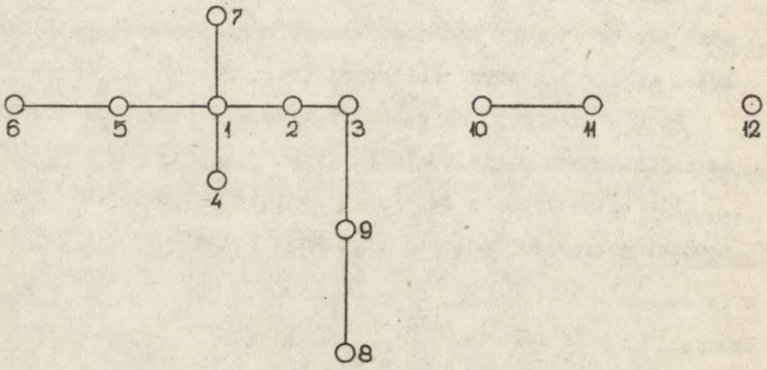
Rys 1. Dendryty rzędu pierwszego



Rys. 2. Najkrótszy dendryt



Rys. 3. Podział najkrótszego dendrytu



http://rcin.org.pl

1	Starogard	5	Kutno	9	Dębica
2	Skierniewice	6	Inowrocław	10	Sanok
3	Jędrzejów	7	Słupsk	11	Lesko
4	Bydgoszcz	8	Końskie	12	Ustrzyki

Podobieństwo między powiatami określono za pomocą odległości euklidesowej obliczanej na podstawie wartości unormowanych (patrz wzór 12). Odległości te podano w tablicy 1.

Na rysunku 1 przedstawiono połączenie rzędu pierwszego, a na rysunku 2 najkrótszy dendryt. Na rysunku 3 pokazano podział dendrytu na 2 i 3 części. Związane z tym podziały na 2 i 3 grupy jak i podziały na więcej grup wynikające z dalszego podziału dendrytu przedstawiono w tablicy 2.

2. MIARY PODOBIEŃSTWA

2.1 Uwagi wstępne

Przeważająca większość metod klasyfikacji wymaga liczbowego określenia stopnia podobieństwa między każdą parą rozpatrywanych obiektów. Miarą podobieństwa będziemy nazywać funkcję $\mu(A_k, A_l)$, która każdej parze obiektów przypisuje wartość liczbową, określającą stopień podobieństwa.

W większości przypadków miarę podobieństwa oblicza się na podstawie cech obserwowanych u badanych obiektów. Jest zupełnie naturalne, by dwa obiekty, które różnią się niewiele pod względem rozpatrywanych cech uznać za podobne, natomiast obiekty różniące się bardzo wartościami cech za niepodobne.

W zależności od sposobu określenia różnicy między wartościami

http://rcin.org.pl

cech otrzymuje się różne funkcje μ (A_k, A_l), a więc różne miary podobieństwa.

O stopniu podobieństwa lub niepodobieństwa będzie oczywiście decydował także dobór cech branych pod uwagę.

Miary podobieństwa podzielić można na dwie grupy. Do pierwszej zalicza się tzw. wskaźniki podobieństwa, do drugiej odległości. Wskaźniki podobieństwa oznaczać będziemy przez $s(A_k, A_l)$ lub s_{kl}, odległości przez $d(A_k, A_l)$ lub d_{kl}.

Duże wartości wskaźnika podobieństwa świadczą o dużym podobieństwie, małe o małym podobieństwie. Przy odległości między obiektami jest na odwrót: małe odległości świadczą o dużym podobieństwie, duże odległości o małym podobieństwie.

2.2. Właściwości wskaźników podobieństwa

Aby używany wskaźnik podobieństwa był sensowny, powinien on spełniać następujące dwa warunki:

Pierwszy, to warunek symetrii

$$s_{kl} = s_{lk} \tag{1}$$

Jest to naturalny warunek, by wielkość podobieństwa nie zależała od kolejności obu obiektów.

Drugi warunek dotyczy unormowania wskaźnika podobieństwa. Żąda się, by był on nieujemny.

$$s_{kl} \geqslant 0 \tag{2}$$

Ponadto pożądane ze względów praktycznych jest, by ustalona była wartość odpowiadająca maksymalnemu podobieństwu, tzn. podobieństwu obiektów identycznych pod względem rozpatrywanych cech. Zwykle wskaźniki podobieństwa konstruuje się tak, by wartość ta wynosiła jeden.

http://rcin.org.pl

Mamy wtedy:

$$s_{kk} = 1 \tag{3}$$

gdzie przez s_{kk} będziemy rozumieć wskaźnik mierzący podobieństwo dowolnego obiektu do samego siebie.

W konsekwencji

$$0 \leq s_{kl} \leq 1 \tag{4}$$

2.3. Przykłady wskaźników podobieństwa

W części tej przedstawimy przykłady wskaźników podobieństwa w zależności od rodzaju cech obserwowanych u badanych osobników.

2.3.1 Wskaźniki podobieństwa w przypadku cech alternatywnych

Przypuśćmy, że u każdego osobnika obserwujemy występowanie lub niewystępowanie p określonych właściwości $X_1, X_2, \ldots, X_p$. Występowanie właściwości X_i oznaczać będziemy przez 1, niewystępowanie przez 0. Wyniki obserwacji można zestawić w tablicy, której fragment przytoczony jest poniżej (ilość cech w tym przykładzie wynosi 10).

W nagłówku tablicy wypisano symbole 10 badanych cech, w kolumnie tytułowej obiekty.

http://rcin.org.pl

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
A_1										
A_2										
.										
.										
.										
A_k	0	0	1	1	0	1	1	1	0	1
.										
.										
.										
A_1	1	0	0	1	0	0	1	1	1	0
.										
.										
.										
.										
A_n										

Następnie dla każdej pary obiektów sporządzić można tablicę czteropolową, w której wypisuje się informacje dotyczące ilości wsółwystępowania zer i jedynek. Tablica ma wygląd jak poniżej:

Obiekt A_k	Obiekt A_1	
	1	0
1	n_{11}	n_{10}
0	n_{01}	n_{00}

W lewej górnej klatce wpisuje się liczbę właściwości występujących jednocześnie u obu obiektów (n_{11}). W prawym dolnym rogu wpisuje się liczbę właściwości niewystępujących ani u jednego ani u drugiego obiektu (n_{00}). Liczba n_{01} to ilość właściwości, które u

http://rcin.org.pl

obiektu drugiego występują a u pierwszego nie, natomiast n_{10} to liczba właściwości, które występują u obiektu pierwszego a u drugiego nie.

Dwa obiekty są tym bardziej podobne, im większe są wartości n_{11} i n_{00}.

Dane zawarte w tablicy czteropolowej są podstawą do obliczenia różnego rodzaju wskaźników podobieństwa.

W przypadku, gdy ilość p rozpatrywanych właściwości, których występowanie się obserwuje, nie może być z góry określona, wtedy liczba n_{00} jest nieznana i prawej dolnej kratki tablicy czteropolowej nie wypełnia się. W sytuacji takiej trzeba stosować takie wskaźniki podobieństwa, których obliczenie nie wymaga znajomości n_{00}.

W dalszym ciągu równoczesne występowanie lub niewystępowanie określonej właściwości u dwóch obiektów nazywać będziemy przypadkiem zgodnym. Ilość przypadków zgodnych wynosi więc

$$z = n_{11} + n_{00}$$

Natomiast występowanie określonej właściwości u jednego obiektu i niewystępowanie u drugiego nazywać będziemy przypadkiem niezgodnym. Ilość przypadków niezgodnych jest więc równa

$$n_{01} + n_{10}$$

Przykłady wskaźników podobieństwa

(Dane bibliograficzne do cytowanych wskaźników znaleźć można w Sokal i Sneath, 1963).

Wskaźnik Sokala i Michenera. Wskaźnik ten jest określony wzorem

$$s_{SM} = \frac{z}{p} = \frac{n_{11} + n_{00}}{n_{11} + n_{01} + n_{10} + n_{00}}$$

http://rcin.org.pl

Jest on tym większy, im większa jest liczba przypadków zgodnych. Maksymalna wartość wskaźnika wynosi jeden i jest osiągnięta, gdy wszystkie właściwości zgodnie występują lub nie występują. Minimalna wartość wynosi zero. Osiąga się ją wtedy, gdy we wszystkich cechach występują niezgodności.

Wskaźnikiem, który nie uwzględnia zgodnego niewystępowania badanych właściwości jest wskaźnik Jaccarda.

$$s_J = \frac{n_{11}}{n_{11} + n_{01} + n_{10}}$$

Podobnie jak poprzednio, maksymalna wartość tego wskaźnika wynosi 1. Wartość ta jest osiągana w przypadku pełnej zgodności. Minimalna wartość wynosi zero i jest osiągana, gdy nie ma przypadku zgodnego współwystępowania.

Podobnym wskaźnikiem jest wskaźnik Russella i Rao

$$s_{RR} = \frac{n_{11}}{p}$$

W odróżnieniu od wskaźnika Jaccarda, wskaźnik ten staje się równy jeden, tylko wtedy, gdy u obu obiektów występują równocześnie wszystkie właściwości.

Oba ostatnie wskaźniki mniejszą wagę przypisują zgodnemu niewystępowaniu badanych właściwości. Można to stwierdzić na następującym przykładzie:

http://rcin.org.pl

A_k : 1 1 1 1 0 0 1 1 0 0 0 0

A_l : 1 1 1 1 1 1 0 0 1 1 0 0

$s_J = \frac{4}{8} = \frac{1}{2} = 0,5$, $s_{RR} = \frac{4}{12} = 0,33$, $s_{SM} = \frac{8}{12}$

A_r : 1 1 1 1 0 0 1 1 1 1 0 0

A_s : 1 1 1 1 1 1 0 0 1 1 0 0

$s_J = \frac{6}{10} = 0,6$, $s_{RR} = \frac{6}{12} = 0,50$, $s_{SM} = \frac{8}{12}$

W pierwszej i drugiej parze obiektów ilość przypadków zgodnych jest taka sama. Dla drugiej jednak pary większa jest ilość przypadków współwystępowania. Wskaźnik Michenera i Sokala jest dla obu par taki sam, natomiast wskaźnik Jaccarda i wskaźnik Russella-Rao są większe dla pary drugiej.

Przypisywanie większej wagi współwystępowaniu ma sens wtedy, gdy samo występowanie określonej właściwości ma większe znaczenie niż jej niewystępowanie.

Wymienione trzy wskaźniki taką samą ważność przypisują przypadkom zgodnym, co niezgodnym. Z istoty niektórych zagadnień może wynikać, że należałoby wyżej oceniać przypadki zgodne lub na odwrót. Następne dwa wskaźniki oceniają z dwukrotnie większą ważnością przypadki zgodne.

Pierwszy z nich to wskaźnik

$$s_1 = \frac{2(n_{11} + n_{00})}{p + n_{11} + n_{00}}$$

Przykład:

A_k : 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0

A_l : 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0

http://rcin.org.pl

$$s_1 = \frac{2 \times 8}{16 + 8} = \frac{16}{24} = \frac{2}{3} = 0{,}67$$

$$s_{SM} = \frac{8}{16} = 0{,}50$$

Podobnym wskaźnikiem, który jednak nie uwzględnia zgodnego niewystępowania jest wskaźnik Dice'a

$$s_D = \frac{2\, n_{11}}{(2\, n_{11} + n_{01} + n_{10})}$$

Dla powyższego przykładu

$$s_D = \frac{2 \times 4}{2 \times 4 + 4 + 4} + \frac{8}{16} = 0{,}50$$

$$s_J = \frac{4}{12} = \frac{1}{3} = 0{,}33$$

Drugim wskaźnikiem jest stosunek ilości zgodnych przypadków do niezgodnych

$$s_2 = \frac{n_{11} + n_{00}}{n_{10} + n_{01}}.$$

Wskaźnik ten przyjmuje wartość zero, gdy nie ma przypadków zgodnych. Natomiast gdy nie ma przypadków niezgodnych jego wartość dąży do nieskończoności.

Podobny jest wskaźnik Kulczyńskiego

$$s_{K1} = \frac{n_{11}}{n_{01} + n_{10}}$$

będący ilorazem ilości współwystępowań do ilości przypadków niezgodnych.

W następnych czterech wskaźnikach występują w mianownikach ilości odnoszące się do poszczególnych cech.

Są to wskaźniki:

$$s_3 = \frac{1}{4}\left[\frac{n_{11}}{n_{11} + n_{10}} + \frac{n_{11}}{n_{11} + n_{01}} + \frac{n_{00}}{n_{01} + n_{00}} + \frac{n_{00}}{n_{10} + n_{00}}\right].$$

http://rcin.org.pl

wskaźnik Kulczyńskiego

$$s_{K2} = \frac{1}{2}\left[\frac{n_{11}}{n_{11} + n_{10}} + \frac{n_{11}}{n_{11} + n_{01}}\right]$$

(wskaźnik podobieństwa systematycznego),

$$s_4 = \frac{n_{11}\ n_{00}}{\sqrt{(n_{11} + n_{10})\ (n_{11} + n_{01})\ (n_{01} + n_{00})\ (n_{10} + n_{00})}}$$

i wskaźnik Ochiai

$$s_O = \frac{n_{11}}{\sqrt{(n_{11} + n_{01})\ (n_{11} + n_{10})}}$$

Wskaźnik s_{K2} Kulczyńskiego i wskaźnik Ochiai nie uwzględniają przypadków wspólnego niewystępowania.

Maksymalna wartość każdego z tych czterech wskaźników jest równa 1. Wartość tę przyjmują wtedy, gdy nie ma przypadków niezgodnych. Wymienione wskaźniki traktują z jednakową ważnością przypadki zgodne co i przypadki niezgodne.

W wymienionych poprzednio wskaźnikach w liczniku nie występowała liczba przypadków niezgodnych. Nie jest tak w następujących wskaźnikach:

Współczynnik Yule'a

$$s_Y = \frac{n_{11}\ n_{00} - n_{01}\ n_{10}}{n_{11}\ n_{00} + n_{01}\ n_{10}}$$

Maksymalna wartość tego wskaźnika wynosząca 1 jest przyjęta wtedy, gdy n_{10} lub n_{00} jest równa zero.

Minimalna wartość wskaźnika wynosi -1. Zachodzi ona, gdy $n_{11} = 0$ lub $n_{00} = 0$.

Ostatni w końcu wskaźnik to współczynnik φ Pearsona

http://rcin.org.pl

$$\varphi = \frac{n_{11}\, n_{00} - n_{01}\, n_{10}}{\sqrt{(n_{11} + n_{10})\,(n_{11} + n_{01})\,(n_{00} + n_{10})\,(n_{00} + n_{01})}}$$

Maksymalna wartość tego wskaźnika jest równa +1, minimalna -1.

2.3.2 Wskaźniki podobieństwa w przypadku cech wielodzielczych

Dotychczas przedstawione wskaźniki odnosiły się do cech alternatywnych, czyli cech polegających na występowaniu określonych właściwości lub objawów. Przejdziemy teraz do przykładów wskaźników podobieństwa, gdy niektóre z badanych cech mogą przyjąć wartości jednej z kilku kategorii.

Tak np. jedna z cech może dzielić się na 2 kategorie, inna na 3, a jeszcze inna na inną liczbę kategorii. Poniżej wypisano fikcyjny przykład.

cechy	A				B		C			D				
kategorie	A_1	A_2	A_3	A_4	B_1	B_2	C_1	C_2	C_3	D_1	D_2	D_3	D_4	D_5
A_k	0	1	0	0	1	0	0	0	1	1	0	0	0	0
A_l	0	1	0	0	0	1	0	0	1	0	0	1	0	0

Jedynkami oznaczono występujące kategorie, zerami niewystępujące.

Jednym z wskaźników podobieństwa dla takich cech jest wskaźnik Rogersa i Tanimoto. Wskaźnik ten jest określony wzorem:

$$s_{RT} = \frac{n_{11}}{n_{01} + n_{10} + n_{11}}$$

gdzie n_{11} jest liczbą kategorii, które występują równocześnie u obu obiektów, natomiast n_{01} i n_{10} są równe ilościom przypadków niezgodnych.

Dla powyższego przykładu wskaźnik Rogersa-Tanimoto jest równy

http://rcin.org.pl

$s_{RT} = \frac{2}{6} = 0,33$

W przypadku, gdy cechy jakościowe mogą być wyskalowane, można podobieństwo oceniać za pomocą wzorów dla cech wielkościowych.

2.4 Miary podobieństwa dla cech wielkościowych

Przypuśćmy teraz, że każdy z badanych obiektów scharakteryzowany jest p cechami wielkościowymi $X_1, X_2, \ldots, X_p$. Wartości tych cech u k-tego osobnika oznaczać będziemy przez

$x_k = (x_{1k}, x_{2k}, \ldots, x_{pk})$.

Jest zupełnie naturalne, by dwa obiekty różniące się niewiele w wartościach badanych cech uznać za podobne, natomiast obiekty różniące się - za niepodobne.

W przypadku cech ciągłych jako miar podobieństwa używa się funkcji charakteryzujących różnicę między badanymi cechami. Funkcje te nazywa się odległościami. Będziemy je oznaczać przez $d(A_k, A_l) = d(x_k, x_l) = d_{kl}$. Im większa odległość, tym mniejsze podobieństwo i na odwrót, im mniejsza odległość, tym większe podobieństwo.

W zależności od tego w jaki sposób mierzyć się będzie różnicę między wartościami cech obu obiektów, otrzymać można różnice miary w odległości.

Przykładem jednej z nich jest tzw. przeciętna różnica (Czekanowski, 1932), określona wzorem

$$d_{1,kl} = \frac{1}{p} \left(|x_{1k} - x_{11}| + |x_{2k} - x_{21}| + \ldots + |x_{pk} - x_{p1}| \right) \qquad (5)$$

gdzie $|x_{ik} - x_{il}|$ oznacza bezwzględną wartość różnicy wartości i-tej cechy.

Inną miarą odległości jest odległość euklidesowa określona

http://rcin.org.pl

wzorem

$$d_{2,kl} = \sqrt{(x_{1k} - x_{11})^2 + (x_{2k} - x_{21})^2 + \ldots + (x_{pk} - x_{p1})^2} \qquad (6)$$

W przypadku dwóch cech odległość ta pokrywa się ze znaną z geometrii płaskiej odległością dwóch punktów o współrzędnych

$P_k = (x_{1k}, x_{2k})$ i $P_l = (x_{11}, x_{21})$

Tak samo w przypadku trzech cech wzór ten określa odległość dwóch punktów w przestrzeni (trójwymiarowej).

Wymaga się, by używana odległość była:

1. nieujemna dla wszystkich par obiektów

$$d_{kl} \geqslant 0, \qquad (7)$$

2. symetryczna, tzn. by jej wartość nie zależała od kolejności obiektów, czyli by

$$d_{kl} = d_{lk}. \qquad (8)$$

3. równa zeru wtedy i tylko wtedy, gdy u obu obiektów wartości cech są takie same

$$d_{kl} = 0 \Longleftrightarrow x_{ik} = x_{il} \quad \text{dla } i = 1, \ldots, p. \qquad (9)$$

Zarówno przeciętna różnica jak i odległość euklidesowa spełniają wszystkie trzy warunki.

Dalszą pożądaną właściwością odległości jest spełnienie prawa trójkąta:

$$d_{kl} \leqslant d_{km} + d_{ml} \qquad (10)$$

Prawo trójkąta oznacza, że bezpośrednia odległość obu obiektów A_k i A_m jest nie większa niż suma odległości obu obiektów od trzeciego. Pojęcie odległości euklidesowej, jakie znamy z geometrii płaskiej i przestrzennej, spełnia prawo trójkąta. Odległość przeciętna spełnia je także.

http://rcin.org.pl

W obu przytoczonych poprzednio wzorach na odległość występowały bezpośrednie (surowe) wartości cech. Ma to tę niedogodność, że odległości między obiektami będą zależały od jednostek, w jakich wyrażone są cechy.

Przypuśćmy na przykład, że chcemy określić podobieństwo ludzi pod względem wysokości i ciężaru ciała za pomocą przeciętnej odległości. Niech jedna osoba ma wysokość ciała 176 cm i ciężar ciała 71 kg, a druga odpowiednio 172 cm i 76 kg. Przeciętna różnica dla tych osób wynosi

$$d_1 = \frac{1}{2}\left\{|176 - 172| + |71 - 76|\right\} = \frac{1}{2}\left\{4 + 5\right\} = 4{,}5$$

Jeżeli natomiast wysokość ciała wyrazić w metrach, to otrzymamy

$$d_1 = \frac{1}{2}\left\{|1{,}76 - 1{,}72| + |71 - 76|\right\} = \frac{1}{2}\left\{0{,}04 + 5\right\} = \frac{5{,}04}{2} = 2{,}52$$

Łatwo sprawdzić, że w drugim przypadku różnica w wysokości ciała prawie nie wpływa na wielkość odległości. Widać stąd, że zwykle wartości badanych cech wymagają odpowiedniego unormowania przed obliczeniem odległości.

Najczęściej stosowane unormowanie cech to odjęcie od wartości cechy jej wartości średniej i podzielenie przez standardowe odchylenie. Unormowane wartości cechy oblicza się wtedy wg wzoru

$$u_i = \frac{x_i - \bar{x}_i}{s_i}, \qquad i = 1, \ldots, p,$$

gdzie i jest numerem cechy, $\bar{x}_i$ średnią arytmetyczną a s_i standardowym odchyleniem.

Wzór na przeciętną różnicę obliczoną z unormowanych wartości cech ma postać:

$$d^*_{1,k1} = \frac{1}{2}\left\{|u_{1k} - u_{11}| + |u_{2k} - u_{21}| + \ldots + |u_{pk} - u_{p1}|\right\} \qquad (11)$$

a wzór na euklidesową odległość

http://rcin.org.pl

$$d^*_{2,kl} = \sqrt{(u_{1k} - u_{11})^2 + (u_{2k} - u_{21})^2 + \ldots + (u_{pk} - u_{pl})^2} \tag{12}$$

Oba te wzory można napisać w postaci zawierającej wyjściowe wartości cechy następująco:

$$d^*_{1,kl} = \frac{1}{p}\left\{\frac{|x_{1k} - x_{11}|}{s_1} + \frac{|x_{2k} - x_{21}|}{s_2} + \ldots + \frac{|x_{pk} - x_{pl}|}{s_p}\right\} \tag{11a}$$

oraz

$$d^*_{2,kl} = \sqrt{\frac{(x_{1k} - x_{11})^2}{s_1^2} + \frac{(x_{2k} - x_{21})^2}{s_2^2} + \ldots + \frac{(x_{pk} - x_{pl})^2}{s_p^2}} \tag{12a}$$

Ponieważ obie te odległości obliczane są z wartości normowanych, a te są bezwymiarowe, więc takie obie odległości nie zależą od jednostek pomiarowych.

Jest to dalsza pożądana właściwość odległości.

Przykład:

W tablicy 3 podano wartości 19 cech charakteryzujących produkcję rolną powiatów augustowskiego, bielsko-podlaskiego i rybnickiego, średnie arytmetyczne i standardowe odchylenia tych cech obliczone z danych dla całej Polski oraz wartości unormowane tych cech dla wymienionych trzech powiatów.

Na podstawie tych wartości unormowanych obliczono odległości między tymi trzema powiatami wg wzoru (12).

Otrzymuje się następujące wyniki:

Augustów - Bielsk Podlaski

$$d^*_2 = \sqrt{\left[0{,}73 - 0{,}55\right]^2 + \left[1{,}98 - 0{,}91\right]^2 + \ldots + \left[-1{,}40 - 0{,}95\right]^2} = 4{,}29$$

http://rcin.org.pl

Augustów - Rybnik

$$d_2^* = \sqrt{\left[0{,}73 - (-2{,}79)\right]^2 + \left[1{,}98 - (-2{,}24)\right]^2 + \ldots + \left[-1{,}40 - 0{,}41\right]^2} =$$

$$= 11{,}33$$

Bielsk Podlaski - Rybnik

$$d_2^* = \sqrt{\left[0{,}55 - (-2{,}79)\right]^2 + \left[0{,}91 - (-2{,}24)\right]^2 + \ldots + \left[-0{,}95 - 0{,}41\right]^2} =$$

$$= 10{,}46$$

Odległości niezależne od jednostek pomiarowych otrzymuje się także wtedy, gdy wartości i-tej cechy podzielić przez rozstęp tej cechy, tzn. przez różnicę między największą i najmniejszą jej wartością.

Jeżeli rostęp i-tej cechy oznaczamy przez R_1, to wzory (9) i (10) przyjmą postać

$$d_1^{**} = \frac{1}{p}\left(\frac{|x_{1k} - x_{11}|}{R_1} + \frac{|x_{2k} - x_{21}|}{R_2} + \ldots + \frac{|x_{pk} - x_{p1}|}{R_p}\right)$$

i

$$d_2^{**} = \sqrt{\frac{(x_{1k} - x_{11})^2}{R_1^2} + \frac{(x_{2k} - x_{21})^2}{R_p^2} + \ldots + \frac{(x_{pk} - x_{p1})^2}{R_p^2}}$$

(We wzorach tych opuszczono przy d^{**} indeksy k, l oznaczające numery obiektów).

Okazuje się, że odległość d_1^{**} jest w badanej grupie obiektów ograniczona od góry przez jedynkę, tzn. że

$$d_1^{**} \leqslant 1$$

dla wszystkich badanych par.

Wszystkie trzy odległości d_2, d_2^* i d_2^{**} nie maleją a zwykle zwiększają się, gdy dodać nowe cechy. Jest to właściwość pożądana; czasem jednak odnosi się obliczoną tymi wzorami odległość do ilości cech, czyli rozpatruje się jako miarę odległości funkcję

http://rcin.org.pl

$$\tilde{d}_2 = \frac{1}{p} d_2$$

i analogicznie dla pozostałych dwóch wzorów.

Jako miary podobieństwa używa się także kwadratu odległości euklidesowej.

Miarą odległości jest także wprowadzony przez Clarka (1952) współczynnik dywergencji

$$d_3 = \sqrt{\frac{1}{p}\left[\left(\frac{x_{1k} - x_{11}}{x_{1k} + x_{11}}\right)^2 + \left(\frac{x_{2k} - x_{21}}{x_{2k} + x_{21}}\right)^2 + \ldots + \left(\frac{x_{pk} - x_{pl}}{x_{pk} + x_{pl}}\right)^2\right]}$$

We współczynniku tym różnica w wartościach poszczególnych cech jest odnoszona do ich sumy.

Różnicę między dwoma obiektami, mierzoną za pomocą odległości euklidesowej obliczanej na podstawie wartości unormowanych cech, można rozbić na składnik mierzący różnicę w średniej wielkości cech oraz na składnik mierzący zróżnicowanie w "profilu" cech.

Weźmy pod uwagę kwadrat odległości euklidesowej podzielonej przez ilość cech $\tilde{d}^2_{2,kl}$. Niech odległość ta będzie liczona dla wartości unormowanych. Ten kwadrat odległości można rozbić na dwa składniki w sposób następujący:

$$\tilde{d}^2_{2,kl} = (\bar{u}_{\cdot k} - u_{\cdot l})^2 + \frac{1}{p}\sum_{i=1}^{p}\left[(u_{ik} - u_{il}) - (\bar{u}_{\cdot k} - \bar{u}_{\cdot l})\right]^2$$

Pierwszy ze składników równy jest kwadratowi różnicy między średnią wielkością wszystkich cech u jednego i drugiego obiektu. Mierzy on więc różnicę między średnią wielkością cech u obu obiektów. Penrose (1954) nazywa go "wskaźnikiem wielkości". Drugi składnik jest równy średniemu kwadratowemu odchyleniu różnic wartości cech obu osobników. Jest on równy zeru wtedy, gdy odchylenie poszczególnych cech od ich wartości średniej dla danego obiektu są równe

http://rcin.org.pl

u obu obiektów. Takie dwa obiekty uważa się za podobne. Drugi wskaźnik nosi nazwę "wskaźnika kształtu". Wskaźniki te wiążą się blisko z pojęciem wskaźników przyrodniczych (Perkal, 1960).

Podejście to pozwala rozbić podobieństwo dwóch obiektów na podobieństwo pod względem abstrakcyjnie rozumianej wielkości oraz podobieństwo także abstrakcyjnie rozumianego kształtu.

Wymienione przykłady wzorów na obliczanie odległości między obiektami nie brały pod uwagę ewentualnej współzależności między cechami. Jeżeli charakteryzujemy obiekty cechami, z których dwie np. są wysoce skorelowane, to duże różnice w jednej cesze będą szły w parze z dużymi różnicami w drugiej cesze, a małe różnice z małymi. Można stać na stanowisku, że traktowanie w takiej sytuacji dwu cech z jednakową wagą pociąga za sobą niepożądany zwiększony wpływ na miarę podobieństwa tego zjawiska, które skorelowane cechy mierzą.

Miarą podobieństwa, która uwzględnia korelację między cechami jest Mahalanobisa odległość uogólniona.

Idea konstrukcji tej odległości jest następująca. Zespół badanych cech przekształca się za pomocą przekształcenia liniowego do unormowanego zespołu zmiennych nieskorelowanych, np.: $Y_1, Y_2, \ldots, Y_p$. Na podstawie tych samych zmiennych oblicza się zwykłą odległość euklidesową. Za pomocą wyjściowych zmiennych odległość uogólnioną można wyrazić wzorem

$$\Delta_{kl} = \sum_i \sum_j (x_{ik} - \bar{x}_{i.})(x_{jl} - \bar{x}_{j.}) c^{ij}$$

gdzie

c^{ij} są wyrazami macierzy odwrotnej do macierzy wariancji-kowariancji.

Wszystkie wymienione dotychczas miary odległości opierały się

http://rcin.org.pl

na wielkościach badanych cech. Odległością, którą oblicza się na podstawie rang poszczególnych cech, jest tzw. odległość rangowa Kendalla.

Oznaczamy przez $R_k^{(i)}$ rangę k-tego osobnika w i-tej cesze. Jest to innymi słowy kolejny numer, jaki ma wartość i-tej cechy u k-tego obiektu w uporządkowaniu wg wielkości. Jeżeli wszystkie rangi są różne, to odległość rangowa określona jest wzorem

$$d_{r,kl} = \frac{3}{p(p^2 - 1)} \sum_i (R_k^{(i)} - R_l^{(i)})^2$$

Odległość ta jest unormowana, tzn.

$$0 \leqslant d_{r,kl} \leqslant 1$$

Co ważniejsze, miara ta jest niezmiennicza ze względu na zmianę jednostek pomiarowych.

Można także w przypadku cech wielkościowych obliczać miary typu wskaźnika podobieństwa. Najczęściej używanym typem takiej miary jest tzw. współczynnik korelacji między obiektami, określony wzorem

$$r_{kl} = \frac{\sum_{i=1}^{p} (x_{ik} - \bar{x}_{\cdot k})(x_{il} - \bar{x}_{\cdot l})}{\sqrt{\sum_{i=1}^{p} (x_{ik} - \bar{x}_{\cdot k})^2 \sum_{i=1}^{p} (x_{il} - \bar{x}_{\cdot l})^2}}$$

gdzie $\bar{x}_{\cdot k}$ i $\bar{x}_{\cdot l}$ są średnimi wartościami cech odpowiednio u obiektów A_k i A_l

$$\bar{x}_{\cdot k} = \frac{1}{p} \sum_{i=1}^{p} x_{ik} \qquad \text{oraz}$$

http://rcin.org.pl

$$\bar{x}_{\cdot 1} = \frac{1}{p} \sum_{i=1}^{p} x_{i1}$$

Współczynnik korelacji między obiektami zawarty jest w granicach -1 i +1.

Maksymalną wartość przyjmuje się wtedy, gdy oba obiekty mają te same wartości cech.

Wskaźnik korelacji między obiektami zależy od jednostek w jakich wyrażono cechy, a nawet od początków skali odczytu.

2.5. Związki formalne między wskaźnikami podobieństwa a odległościami

Miary podobieństwa typu odległość można za pomocą odpowiednich wzorów przekształcić na wskaźniki podobieństwa i odwrotnie. Oznaczamy przez d_{kl} odległość między obiektami A_k i A_l, a przez s_{kl} ich wskaźnik podobieństwa.

Załóżmy, że $0 \leqslant s_{kl} \leqslant 1$ oraz $d_{kl} \geqslant 0$.

Wtedy wzorami takimi są:

$$d_{kl} = 1 - s_{kl}$$

$$d_{kl} = \sqrt{1 - s_{kl}}$$

$$d_{kl} = - \log s_{kl}$$

$$s_{kl} = 1 - \frac{d_{kl}}{\max d_{kl}}$$

$$s_{kl} = 1 - \left(\frac{d_{kl}}{\max d_{kl}} \right)^2$$

Czasem obiektami porządkowanymi są same cechy. W przypadku cech wielkościowych używa się jako miary podobieństwa współczynnika korelacji r między nimi, określonego wzorem

http://rcin.org.pl

$$r_{ij} = \frac{\sum_{k=1}^{n} (x_{ik} - \bar{x}_{i.})(x_{jk} - \bar{x}_{j.})}{\sqrt{\sum_{k=1}^{n} (x_{ik} - \bar{x}_{i.})^2 \sum_{k=1}^{n} (x_{jk} - \bar{x}_{j.})^2}}$$

gdzie $\bar{x}_{i.}$ i $\bar{x}_{j.}$ są średnimi wartościami obu cech.
Współczynnik korelacji zawiera się w granicach -1 i +1. Używanie go jako miary podobieństwa ma sens, jeżeli wszystkie cechy są skorelowane nieujemnie.

3. KRYTERIA OPTYMALNOŚCI PODZIAŁÓW

3.1 Wstęp

Celem klasyfikacji jest podział zbioru obiektów na grupy jednorodne. Wykonanie tego zadania wymaga więc określenia, jak będzie się mierzyć jednorodność grup. Miary te powinny pozwolić na rozstrzygnięcie, w którym z dwóch podziałów grupy są bardziej jednorodne, czyli który z dwóch podziałów uznać należy za lepszy. Kryterium takie określa równo-cześnie, który podział będzie najlepszy.

Przypuśćmy, że zbiór obiektów podzielono na g grup $G_1, G_2, \ldots, G_g$. Ilość obiektów w j-tej oznaczamy przez n_j. Same zaś obiekty należące do tej grupy przez $A_{j1}, A_{j2}, \ldots, A_{jn_j}$. Podział taki oznaczamy ogólnie przez P.

Podział jest w pełni scharakteryzowany przez wymienienie obiektów w poszczególnych grupach, a cząstkowymi jego charakterystykami są ilość grup oraz ilość obiektów w poszczególnych grupach.

http://rcin.org.pl

Przez kryterium dobroci podziału będziemy rozumieli funkcję, która każdemu podziałowi przyporządkowuje liczbę K(P).

Przykład kryterium może być następujący. Niech $\min s_{j,kl}$ oznacza najmniejszy wskaźnik podobieństwa w j-tej grupie. Kryterium może być wtedy suma minimalnych wskaźników podobieństw we wszystkich grupach, czyli funkcja

$$K(P) = \sum_{j=1}^{g} \min s_{j,kl}$$

Jeżeli mamy dwa podziały P_1 i P_2, to za lepszy uznać należy ten, dla którego wartość K(P) jest większa. Czyli P_1 będzie lepsze od P_2, gdy $K(P_1) > K(P_2)$.

Jeżeli jako miary podobieństwa używamy odległości wtedy poprzednie kryterium będzie postaci

$$K(P) = \sum_{j=1}^{g} \max d_{j,kl}$$

gdzie $\max d_{j,kl}$ oznacza największą odległość obiektów w j-tej grupie. W sensie tego kryterium podział będzie tym lepszy, im mniejsza będzie wartość K.

Kryteria dobroci mogą brać pod uwagę jednorodność wewnątrz grup, podobieństwo między grupami lub równocześnie jedno i drugie.

3.2 Miary jednorodności wewnątrz grup

Poniżej podano kilka wskaźników mierzących jednorodność grupy. Minimalny wskaźnik podobieństwa w j-tej grupie

$$m_j = \min s_{j,kl} \qquad (13)$$

Średni wskaźnik podobieństwa w grupie

http://rcin.org.pl

$$\bar{s}_j = \frac{1}{n_j(n_j - 1)} \sum_{k=1}^{n_j} \sum_{l=1}^{n_j} s_{j,kl} \qquad (14)$$

Średnia ważona wskaźników podobieństwa w grupach

$$\bar{S}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} \sum_{l=1}^{n_j} s_{j,kl} \qquad (15)$$

Maksymalna odległość w j-tej grupie

$$D_j = \max d_{j,kl} \qquad (16)$$

Średnia odległość w grupie

$$\bar{d}_j = \frac{1}{n_j(n_j - 1)} \sum_{k=1}^{n_j} \sum_{l=1}^{n_j} d_{j,kl} \qquad (17)$$

Średnia odległość od środka ciężkości.

Przez środek ciężkości grupy obiektów rozumie się fikcyjny obiekt o wartościach cech równych średnim arytmetycznym wartości cech obiektów należących do tej grupy. Jeżeli więc przez x_{ijk} oznaczyć wartość i-tej cechy u k-tego obiektu w j-tej grupie, to wartość i-tej cechy środka ciężkości będzie równa:

$$\bar{x}_{ij\cdot} = \frac{1}{n_j} \sum_{k=1}^{n_j} x_{ijk} \qquad (18)$$

Średni kwadrat odległości w grupie G_j od środka ciężkości wynosi wtedy:

$$\delta_j^2 = \frac{1}{n_j} \sum_{k=1}^{n_j} \left(\sum_{i=1}^{p} (x_{ijk} - \bar{x}_{ij\cdot})^2 \right) \qquad (19)$$

http://rcin.org.pl

3.3 Miary podobieństwa grup

Poniżej podano miary określające podobieństwo dwóch grup. Wskaźniki maksymalnego podobieństwa grup określa się jako maksymalny wskaźnik podobieństwa między obiektami jednej i drugiej grupy:

$$Ms_{uv} = \max_{A_k \in G_u, A_l \in G_v} s_{kl} \tag{20}$$

Wskaźnik minimalnego podobieństwa grup określa się jako najmniejszy wskaźnik podobieństwa między obiektami jednej i drugiej grupy

$$ms_{uv} = \min_{A_k \in G_u, A_l \in G_v} s_{kl} \tag{21}$$

Podobne wskaźniki tworzy się dla odległości:
Minimalna odległość między grupami

$$md_{uv} = \min_{A_k \in G_u, A_l \in G_v} d_{kl} \tag{22}$$

Maksymalna odległość między grupami

$$Md_{uv} = \max_{A_k \in G_u, A_l \in G_v} d_{kl} \tag{23}$$

Średnie podobieństwo między grupami G_u i G_v

$$\bar{s}_{uv} = \frac{1}{n_u n_v} \sum_{A_k \in G_u} \sum_{A_l \in G_v} s_{kl} \tag{24}$$

Kwadrat odległości między środkami ciężkości grup G_u i G_v

$$D^2_{uv} = \sum_{i=1}^{p} (\bar{x}_{i.u} - \bar{x}_{i.v})^2 \tag{25}$$

http://rcin.org.pl

3.4 Kryteria dobroci podziału

W rozdziale tym podane są niektóre kryteria dobroci podziału.

1. Suma minimalnych wskaźników podobieństwa w grupach.

$$K_1(P) = \sum_{j=1}^{g} m_j = \sum_{j=1}^{g} \min s_{j,kl}$$

Według tego kryterium za lepszy z dwóch podziałów na tę samą ilość grup należy uznać ten, dla którego $K_1(P)$ jest większe. Najlepszy podział P_{opt} będzie ten, dla którego K_1 osiągnie wartość maksymalną.

2. Innym kryterium jest minimalna wartość najmniejszych wskaźników podobieństwa.

$$K_2(P) = \min_j \quad m_j = \min_j \quad \min_{k,l} s_{j,kl}$$

Według tego kryterium najlepszym podziałem będzie także ten, dla którego wartość K_2 jest możliwie duża.

3. Suma średnich podobieństw w grupach.

$$K_3(P) = \sum_{j=1}^{g} \bar{s}_j \qquad \text{lub}$$

$$K_4(P) = \sum_{j=1}^{g} \bar{\bar{s}}_j$$

Lepszy z dwóch jest ten podział, dla którego wartość K_3 lub K_4 jest większa.

Podobne kryteria można sformułować dla odległości a więc:

http://rcin.org.pl

4. Suma maksymalnych odległości w grupach

$$K_5(P) = \sum_{j=1}^{g} D_j$$

5. Maksymalna z maksymalnych odległości w grupach

$$K_6(P) = \max_j D_j$$

6. Suma średnich odległości wewnątrz grup

$$K_7(P) = \sum_{j=1}^{g} \bar{d}_j$$

7. Suma średnich odległości od środków ciężkości wewnątrz grup

$$K_8(P) = \sum_{j=1}^{g} \delta_j$$

Ponieważ wszystkie te kryteria obliczane są z odległości, lepszym podziałom będą odpowiadały mniejsze wartości. Podobne kryteria można budować opierając się na miarach podobieństwa między grupami.

Można połączyć kryteria oceny jednorodności wewnątrz grup i kryteria oceny niejednorodności między grupami w jedno łączne kryterium.

Dla wskaźników podobieństwa może ono być postaci:

$$W_1 = \left(\sum_{j=1}^{g} \frac{1}{n_j(n_j-1)} \sum_k \sum_l s_{j,kl} \right) \Bigg/ \left(\sum_{u \neq v} \sum \frac{1}{n_u} \cdot \frac{1}{n_v} \sum_{A_k \in G_u} \sum_{A_l \in G_v} s_{kl} \right).$$

W myśl tego kryterium za lepszy należy uznać ten podział, dla którego W_1 jest większe.

Dla odległości otrzymujemy analogiczne kryterium

$$W_2 = \left(\sum_{j=1}^{g} \frac{1}{n_j(n_j-1)} \sum_k \sum_l d_{j,kl} \right) \Bigg/ \left(\sum_{u \neq v} \sum \frac{1}{n_u} \cdot \frac{1}{n_v} \sum_{A_k \in G_u} \sum_{A_l \in G_v} s_{kl} \right)$$

http://rcin.org.pl

Dość prostej postaci są następujące dwa kryteria.

Niech s^* będzie ustaloną wartością krytyczną wskaźnika podobieństwa. Jako kryterium dobroci podziału używa się funkcji

$$C = \sum_1 (s_{kl} - s^*) - \sum_2 (s_{kl} - s^*) .$$

Symbol $\sum_1$ oznacza sumę dla par obiektów należących do tej samej klasy, $\sum_2$ sumę dla wszystkich par należących do różnych klas. W myśl tego kryterium ten podział należy uznać za lepszy, dla którego C jest większe. Kryterium to można uprościć, jeżeli wprowadzić nową zmienną v_{kl} wg przepisu

$$v_{kl} = 1 \qquad \text{gdy} \quad s_{kl} - s^* \geq 0$$

$$v_{kl} = 0 \qquad \text{gdy} \quad s_{kl} - s^* < 0$$

Otrzymuje się wtedy kryterium

$$V = \sum_1 v_{kl} - \sum_2 v_{kl} ,$$

którego wartość jest równa ilości jedynek wewnątrz klas minus ilość jedynek między klasami.

4. PROCEDURY GRUPOWANIA

4.1 Wstęp

Przez procedurę grupowania będziemy rozumieć postępowanie, które daje podział rozważanego zbioru obiektów na grupy rozłączne.

W niniejszym opracowaniu będziemy zajmować się procedurami, które korzystają z tablic wskaźników podobieństwa lub odległości.

Jeżeli ustalone jest obiektywne ilościowe kryterium optymalności podziału, to teoretycznie optymalny podział może być znaleziony drogą sprawdzenia wszystkich możliwych podziałów i wyboru

http://rcin.org.pl

najlepszego. W praktyce nie można z tych sposobów korzystać, gdyż ilość wszystkich podziałów, nawet przy niewielkiej ilości obiektów, jest tak duża, że zadanie sprawdzenia wszystkich podziałów staje się niemożliwe nawet przy użyciu bardzo szybko liczących komputerów. Dlatego też stosuje się procedury, które szukają rozwiązania optymalnego lub zbliżonego do optymalnego metodą kolejnych przybliżeń.

Wśród procedur grupowania wyróżniają się procedury prowadzące do hierarchii podziałów. Procedury te mogą być łączeniowe lub podziałowe.

Procedury łączeniowe pracują następująco. Punktem wyjściowym jest podział na n grup jednoobiektowych. W kolejnych krokach zmniejsza się ilość grup, łącząc wg ustalonych zasad po dwie grupy w jedną, aż w końcu uzyskuje się jedną grupę obejmującą wszystkie obiekty.

W procedurach podziałowych zaczyna się od jednej grupy obejmującej wszystkie obiekty. W kolejnych etapach zwiększa się ilość grup przez podział jednej z grup na dwie części. W końcu dochodzi się do podziału na n grup jednoobiektowych.

4.2 Przykłady procedur grupowania

4.2.1 Wprowadzenie

W rozdziale tym omówimy najbardziej znane i równocześnie najprostsze procedury grupowania.

Procedury te są ilustrowane na przykładzie 12 powiatów podanych na str. 18. Praktyczne zastosowanie procedur upraszcza się znacznie, gdy skorzystać z odległości uporządkowanych wg wielkości. Dla naszego przykładu odległości te podano w tablicy 4.

Jedną z procedur grupowania, tzw. taksonomię wrocławską, przed-

http://rcin.org.pl

stawiono w rozdziale 2. Jest to procedura podziałowa. Daje ona te same grupy co tzw. metoda najbliższego sąsiada, omówiona także w dalszej części.

Jedną z najwcześniejszych "metod" grupowania a zarazem metodą porządkowania, jest metoda diagramu Czekanowskiego. Jest to wprawdzie metoda niesformalizowana i w tym sensie nieobiektywna, ale jej praktyczna przydatność jest duża.

4.2.2 Metoda diagramu Czekanowskiego

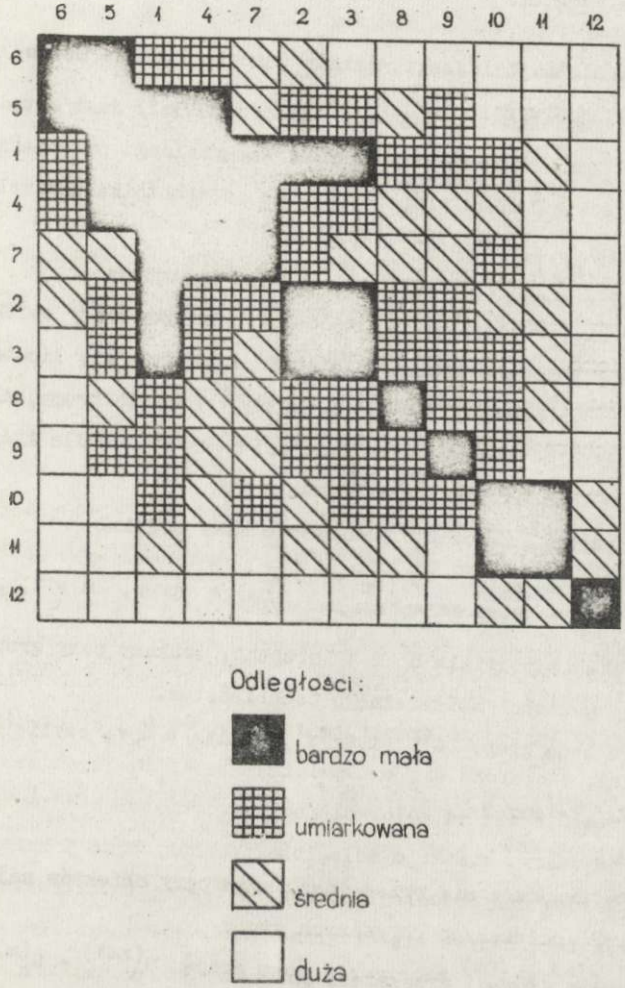
Metoda ta polega na odpowiednim porządkowaniu obiektów za pomocą tablicy odległości lub tablicy wskaźników podobieństwa. Mianowicie przedstawia się obiekty i odpowiadające im wiersze i kolumny tablicy odległości (wskaźniki podobieństwa) tak, by małe odległości (lub duże wskaźniki podobieństwa) skupiały się możliwie blisko głównej przekątnej. Uporządkowanie takie odzwierciedla fakt, że obiekty sąsiednie są bardzo podobne.

Dla ułatwienia całego postępowania i uzyskania plastyczności zastępuje się wyjściową tablicę odległości (wskaźników podobieństwa) diagramem. Diagram sporządza się w ten sposób, że pola tablicy odległości zaczernia się w odpowiedni sposób. Mianowicie odległości występujące w tablicy dzieli się według ich wielkości na kilka kategorii. Następnie kratki, w których występują odległości bardzo małe, zaczernia się całkowicie. Kratki o odległościach umiarkowanych maluje się na kolor ciemnoszary, kratki o odległościach średnich na szary, kratki o odległościach dużych na jaśniejszy, a kratki o bardzo dużych odległościach zostawia się całkowicie białe.

Diagram taki nazywa się diagramem Czekanowskiego. Dążenie, by skupić wzdłuż głównej przekątnej pola o małych odległościach (dużych wskaźników podobieństwa) sprowadza się do uzyskania możliwie

http://rcin.org.pl

Rys.4. Diagram Czekanowskiego



http://rcin.org.pl

dużego zaczernienia wzdłuż głównej przekątnej.

Zwykle na diagramie takim wyodrębniają się ciemne kwadraty, które odzwierciedlają fakt układania się obiektów w grupy.

Na rysunku 4 pokazano diagram Czekanowskiego 12 powiatów podanych na str. 18.

4.2.3 Metoda najbliższego sąsiada (Single linkage method)

Metoda najbliższego sąsiada (Johnson, 1967) jest metodą łączeniową. Korzysta ona ze wskaźnika maksymalnego podobieństwa grup Ms_{uv} określonego wzorem (20) lub ze wskaźnika minimalnej odległości md_{uv} określonego wzorem (22).

Procedura najbliższego sąsiada jest następująca:
Wychodzi się od podziału n-grupowego, w którym każdy obiekt stanowi grupę jednoelementową. Następnie zmniejsza się liczbę grup krok po kroku o jedną, łącząc za każdym razem te grupy, dla których maksymalne podobieństwo jest największe (lub dla których minimalna odległość jest najmniejsza).

Przypuśćmy, że na którymś etapie mamy podział

$$P(m) = \left\{G_1^{(m)}, G_2^{(m)}, \ldots, G_m^{(m)}\right\} \quad \text{na } m \text{ grup, } m = 1, \ldots, n-1.$$

Dla uzyskania podziału o m-1 grupach, szukamy pary grup o największej wartości maksymalnego podobieństwa.
Niech to będą grupy $G_u^{(m)}$ i $G_v^{(m)}$, gdzie $u < v$, czyli

$$Ms_{uv} = \max_{r \neq t} Ms_{rt}$$

Grupy te znajduje się przez odszukanie pary obiektów najbardziej podobnych, należących do różnych grup.
Z grup $G_u^{(m)}$ i $G_v^{(m)}$ tworzy się nową grupę $G_u^{(m-1)} = G_u^{(m)} \cup G_v^{(m)}$.
Wszystkie pozostałe grupy pozostają bez zmian.

http://rcin.org.pl

W przypadku, gdy korzysta się z tablicy odległości, łączy się parę grup, dla których najbliższa odległość jest najmniejsza, czyli parę grup, dla której

$$md_{uv} = \min_{r \neq t} md_{rt} .$$

Przykład.

W tablicy 5 podano hierarchię podziałów 12 powiatów omówionych w punkcie 1.4.2.

Postępowanie zaczyna się od podziałów na 12 grup jednoobiektowych. Następnie w tablicy uporządkowanych odległości (tabl. 4) znajdujemy najbliższą sobie parę. Tworzą ją powiaty 2 i 3, których odległość wynosi 2.40. Połączenie tych dwóch powiatów w jedną grupę daje podział na 11 grup. Następnie w tablicy 2 znajdujemy następną najbliższą parę grup. Są to powiaty 1 i 4. Łącząc je otrzymuje się podział na 10 grup. Następną najbliższą parą grup są grupy (1,4) i (2,4). Wynika to stąd, że następna co do wielkości jest odległość między powiatami 1 i 2, a jeden z tych powiatów należy do jednej, drugi do drugiej z tych grup. Następną co do wielkości jest odległość między powiatami 1 i 3. Nie uwzględniamy jej, bo i tak powiaty 1 i 3 należą już do tej samej grupy. Kolejna co do wielkości jest odległość między powiatem 1 i powiatem 7. Powiat 7 zostaje połączony z grupą, do której należy powiat pierwszy. Postępując tak dalej otrzymuje się ciąg podziałów w tablicy 7. Odległości, które decydują o połączeniu, są właśnie najmniejszymi odległościami między grupami.

4.2.4 Metoda najdalszego sąsiada (Complete linkage method)

Podobnie jak poprzednio, metoda ta jest metodą łączeniową (Johnson, 1967). Inaczej niż metoda najbliższego sąsiada posługuje

http://rcin.org.pl

się ona wskaźnikiem minimalnego podobieństwa grup ms_{uv} określonego wzorem (21), lub wskaźnikiem maksymalnej odległości Md_{uv} między grupami (23).

Łączenie odbywa się tak samo jak w metodzie najbliższego sąsiada, z tym, że łączy się grupy o największej wartości majmniejszego podobieństwa, lub, w przypadku odległości, grupy o najmniejszej wartości najdalszej odległości. Niech takimi grupami będą $G_u^{(m)}$ i $G_v^{(m)}$, czyli

$$ms_{uv} = \max_{r \neq t} ms_{rt} \qquad \text{lub}$$

$$Md_{uv} = \min_{r \neq t} Md_{rt}.$$

Wtedy tworzy się nową grupę

$$G_u^{(m-1)} = G_u^{(m)} \cup G_v^{(m)}.$$

Pozostałe grupy zostają bez zmian.

Przykład.

Hierarchię podziałów uzyskanych metodą najdalszego sąsiada podano w tablicy 6.

Jak w metodzie poprzedniej zaczyna się od podziału na 12 jednoobiektowych grup. Ponieważ najmniejsze odległości między nimi są równocześnie najdalszymi, podział na n-1 = 11 grup jest taki sam co w metodzie najbliższego sąsiada. Dalsze podziały mogą się różnić, chociaż w naszym przykładzie podział na 10 grup jest identyczny, ponieważ odległość powiatów 1 i 4 jest równocześnie najmniejszą z najdalszych odległości między grupami. Ale już podział na 9 grup jest inny. Mianowicie najdalsze obiekty należące jeden do grupy (1,4), drugi do grupy (2, 3), to powiaty 3 i 4 o odległości 5.69. W tablicy odległości uporządkowanych opuszczamy więc

http://rcin.org.pl

odległości między powiatami 1 i 2, 1 i 3 oraz 1 i 7. Następna co do wielkości jest odległość między powiatem 7 i powiatem 4 należącym do grupy (1, 4), wynosząca 3.81. Odległość ta jest wartością wskaźnika najdalszej odległości grupy złożonej z powiatu 7 i grupy złożonej z powiatów (1, 4). Jest to równocześnie najmniejsza z takich odległości pozostałych grup. Powiat siódmy zostaje więc połączony w jedną grupę z powiatami 1 i 4. Następne dwie odległości $d_{1,5}$ i $d_{4,5}$ opuszczamy. Powiatu 5 nie zaliczamy do grupy, do której należy powiat 1, ponieważ największa odległość powiatu 5 od tej grupy to odległość między powiatami 5 i 7 równa 6.30. Następna to odległość między powiatami 5 i 6, i te dwa powiaty zostają połączone w jedną grupę.

Postępując dalej w ten sposób uzyskuje się podziały podane w tablicy 4.

4.2.5 Metody podziałowe Huberta

Metody te (Hubert, 1973) są także hierarchiczne. W odróżnieniu od metody najbliższego sąsiada zaczyna się postępowanie od jednej grupy obejmującej wszystkie obiekty. Liczbę grup zwiększa się krok po kroku o jedną przez rozdzielenie grupy na dwie według określonej zasady. Istnieją trzy różne wersje zasady podziału, które oznaczone będą przez A, B i C. Pierwszy krok podziału jest wspólny wszystkim trzem wymienionym wersjom.

Przypuśćmy, że w kolejnym kroku uzyskaliśmy m grup $G_1^{(m)}$, $G_2^{(m)}$, ..., $G_m^{(m)}$. Znajdujemy tę grupę, $G_u^{(m)}$, która zawiera najbardziej niepodobne (najdalsze) obiekty. Niech to będą obiekty A_k i A_l. Grupa $G_u^{(m)}$ zawierająca te obiekty będzie grupą dzieloną na dwie nowe, G' i G''. Obiekt A_k zalicza się do jednej z nich, np. do G', obiekt A_l do drugiej, G''. Rozdzielanie dalszych obiektów

http://rcin.org.pl

grupy G_u do G' lub G'' odbywa się już według zasad różnych dla wersji A, B i C.

Wersja A.

Wśród obiektów grupy G_u szuka się obiektu A_t najbardziej niepodobnego do obiektu A_k lub A_l. Jeżeli obiekt A_t jest najbardziej niepodobny do A_k zalicza się go do grupy G'', jeżeli jest najbardziej niepodobny do A_l zalicza się go do grupy G'. Postępowanie kontynuuje się wg następującej zasady. Za każdym razem szuka się obiektu najbardziej niepodobnego do obiektów zaliczanych już do grupy G' lub G'' spośród pozostałych jeszcze obiektów grupy G_u. Niech parą takich obiektów będzie A_r i A_s. Wtedy, jeżeli A_r należy do G', to A_s zalicza się do G'', jeżeli natomiast A_r należy do G'', to A_s zalicza się do G'. Grupy G'' i G' stanowią nowe grupy przy podziale na m + 1 grup. Wszystkie pozostałe grupy pozostają bez zmian.

Przykład.

Hierarchię podziałów uzyskanych tą metodą podano w tablicy 7. Postępowanie zaczyna się od jednej grupy zawierającej wszystkie obiekty.

Najbardziej niepodobne powiaty to powiaty 6 i 12 o odległości d(6,12) = 12.02. Zaliczamy powiat 6 do grupy G' a powiat 12 do grupy G''.

W tablicy odległości (tabl. 1) lub tablicy odległości uporządkowanych (tabl. 4) znajdujemy powiat najdalszy od powiatów 6 lub 12. Jest to powiat 11, który jest bardziej odległy od powiatu 6 niż 12. Powiat ten zostaje zaliczony do grupy G'', do której nie należy grupa 6. Szukamy następnie powiatu, który jest najbardziej odległy od powiatu 6, 11 lub 12. Jest to powiat 9, odległy najbardziej od powiatu 12, d(9,12) = 10.94. Powiat ten zostaje zaliczony do grupy G'. Do tej samej grupy zaliczone zostają następne kolejno

http://rcin.org.pl

powiaty 8, 5, 3, 2 i 1.

Następna najbardziej odległa para powiatów, to powiaty 6 i 10, d(6,10) = 8.99. Powiat 10 zostaje więc zaliczony do grupy G'', czyli do tej grupy, do której powiat 6 nie należy.

Kolejne co do odległości to powiaty 6 i 8. Przypadek ten pomijamy, bo powiat 8 został już wcześniej zaliczony do grupy G'.

W ten sposób otrzymujemy podział wszystkich powiatów na dwie grupy

$$G_1^{(2)} = G' = (6,9,8,5,3,2,1,7,4)$$

$$G_2^{(2)} = G'' = (12,11,10)$$

Należy zwrócić uwagę na fakt, że do grupy $G_1^{(2)}$ należą powiaty 6 i 8 bardziej odległe /d(6,8) = 8.37/ niż powiaty 7 i 12 /d(7,12) = 8.82/ oraz 4 i 12 /d(4,12) = 8.73/ należące do różnych grup.

Żeby zdecydować, która z obu grup będzie w następnym kroku dzielona, trzeba znaleźć w jednej i drugiej grupie parę najbardziej odległych powiatów. W grupie pierwszej są to powiaty 6 i 8 o odległości d(6,8) = 8.97, w grupie drugiej powiaty 10 i 12 o odległości d(10,12) = 7.28. Ponieważ grupa pierwsza jest w tym sensie bardziej niejednorodna, ona właśnie podlega podzieleniu. Oznaczamy grupę, której początkiem jest powiat 6 przez G', grupę, której pierwszym obiektem jest powiat 8 przez G''. Postępując jak poprzednio, ograniczając się jednak do obiektów należących do $G_1^{(2)}$ otrzymujemy następujące rozbicia:

$$G' = (6,7,4,5)$$

$$G'' = (8,9,3,2,1)$$

Wprowadźmy następującą numerację:

$$G_1^{(3)} = G'$$

$$G_2^{(3)} = G''$$

http://rcin.org.pl

$$G_3^{(3)} = G_2^{(2)}$$

Kontynuując opisane postępowanie otrzymuje się kolejne podziały jak w tablicy 5.

Wersja B.

Jak w wersji A dwa najbardziej niepodobne obiekty A_k i A_l są pierwszymi obiektami grup G' i G''. Szuka się teraz obiektu najbardziej podobnego (najbliższego) do A_k lub A_l. Niech obiektem tym będzie obiekt A_t. Jeżeli obiekt ten jest podobniejszy do A_k, to zalicza się go także do G', jeżeli jest podobniejszy Do A_l zalicza się go do G''. Postępowanie to kontynuuje się według następującej zasady. Za każdym razem szuka się w G_u obiektu najbardziej podobnego do któregoś z obiektów G' lub G''. Niech parą takich obiektów będą A_r i A_t. Jeżeli A_r należy do G' to A_t zalicza się także do G', jeżeli A_r należy do G'', to A_l zalicza się także do G''.

Przykład.

Hierarchię podziałów uzyskaną wersją B metody podziałowej przedstawiono w tablicy 8.

Tak samo jak poprzednio dwa najbardziej niepodobne powiaty, powiat 6 i 8 tworzą początki dwóch nowych grup. Niech powiat 6 będzie początkiem grupy G', powiat 12 początkiem grupy G''.

Następnie szuka się obiektu najbardziej podobnego do 6 lub 12. Najbliższy jest powiat 5 do powiatu 6. Powiat ten zostaje zaliczony do grupy G'. Szukamy teraz następnego powiatu o najmniejszej odległości od powiatów 5,6 lub 12. Okazuje się, że jest to powiat 1, który jest najbliższy do powiatu 5 /d(1,5) = 4.02/. Powiat ten zostaje zaliczony do G'. Postępując w ten sposób zaliczamy kolejno do grupy G'powiaty 4 /d(1,4) = 3.11/, 2 /d(1,2) = 3.14/, 3 /d(1,3) = 2.40/, 7 /d(1,7) = 3.78/, 9 /d(3,9) = 4.78/,

http://rcin.org.pl

8/d(8,9) = 5.08/, 10 /d(5,10) = 5.25/ i 11 /d(10,11) 5 4.26/. Grupa G' składa się więc z wszystkich powiatów z wyjątkiem 12, który tworzy odrębną grupę.

Oznaczamy $G_1^{(2)} = G'$ i $G_2^{(2)} = G''$. Ponieważ grupa $G_2^{(2)}$ jest jednoelementowa podziałowi podlega następnie $G_1^{(2)}$. Parą najbardziej odległych powiatów w tej grupie są powiaty 6 i 11 i one stanowią początki nowych grup G' i G''. Stosując to samo postępowanie co poprzednio, otrzymuje się następujący podział na grupy

$$G_1^{(3)} = G' = (6,5,1,4,2,3,7,9,8)$$

$$G_2^{(3)} = G'' = (11,10)$$

Trzecia grupa zostaje bez zmian $G_3^{(3)} = G_2^{(2)}$.

Z tych grup najbardziej niepodobną parę powiatów zawiera grupa $G_1^{(3)}$. Są to powiaty 6 i 8. One to tworzą początek dwu nowych grup.

Kontynuowanie postępowania daje takie podziały jak pokazano w tablicy 8.

Wersja C.

Początek podziału grupy G_u na grupy G' i G'' jest taki sam co w wersjach A i B. Zmiana następuje dopiero od przydzielania dalszych obiektów. Zasada jest następująca. Wśród pozostałych jeszcze w grupie G_u obiektów znajduje się obiekt, którego najmniejszy wskaźnik podobieństwa do obiektów zaliczonych już do grupy G' lub G'' jest największy. Niech to ma miejsce dla obiektów A_s i A_t, przy czym A_t jest obiektem pozostałym jeszcze w G_u. Obiekt A_t zostaje wtedy zaliczony do tej samej grupy co A_s.

Przykład.

Podziały uzyskane wersją C metody podziałowej podano w tablicy 9.

http://rcin.org.pl

Początek jest taki sam jak w wersji B. Początkiem grupy G'' jest powiat 12. Podobnie jak poprzednio do grupy G' zaliczony zostaje najpierw powiat 5 jako najbliższy powiatu 6. Następnie szukamy powiatu, którego większa odległość do powiatu 5 i 6 jest najmniejszą. Jest nim powiat 4, / d(4,5) = 4.04, d(4,6) = 4.99 /. Ponieważ nie ma takiego powiatu, którego odległość od powiatu 12 byłaby mniejsza niż 4.99, powiat 4 zaliczony zostaje do grupy G'. Szukamy teraz powiatu, którego największa odległość do powiatów 6, 5 i 4, należących do G', jest możliwie mała. Powiatem tym jest powiat 1. Odległość ta wynosi d(1,6) = 5.74. Ponieważ nie ma także powiatu bliższego 12, powiat 1 zaliczony zostaje do grupy G'. Szukamy ponownie powiatu, którego największa odległość do powiatów grupy G' jest możliwie mała. Jest nim powiat 7 z odległością d(6,7) = 6.63. Ponieważ jednak powiat 11 jest bliższy powiatowi 12, najpierw zaliczamy 11 do grupy G''.

Znajdujemy następnie powiat, którego najdalsza odległość do powiatów G'' jest możliwie mała. Jest nim powiat 10 z odległością d(10,12) = 7.28. Ponieważ największa odległość powiatu 7 od grupy G' jest mniejsza niż najdalsza odległość powiatów 10 od grupy G'', zaliczamy najpierw powiat 7 do grupy G'. Postępując w ten sposób dalej otrzymuje się podział

$$G_1^{(2)} = G' = (6,5,4,1,7,2,2,9,8)$$

$$G_2^{(2)} = G'' = (12,11,10)$$

Dalsze podziały podane są w tablicy 3. W naszym przykładzie wersja C dała te same wyniki co wersja A. Zbieżność ta jest przypadkowa.

http://rcin.org.pl

4.2.6 Metoda łączonych par Sokala i Michenera (pair-group method)

Sokal i Michener /Sokal, Sneath, 1963/ są autorami grupy metod, które pozwalają uwzględnić nie tylko ekstremalne wartości wskaźników podobieństwa (lub odległości), ale wszystkie wskaźniki podobieństwa wszystkich grup.

Jako przykład omówimy metodę łączonych par grup. Metoda ta jest metodą łączeniową. Podobnie jak w metodzie najbliższego sąsiada, łączy się na każdym etapie dwie "najpodobniejsze" grupy w jedną nową. Podczas gdy w metodzie najbliższego sąsiada podobieństwo dwóch grup mierzy się maksymalnym wskaźnikiem podobieństwa, to w metodzie Sokala i Michenera podobieństwo dwóch grup mierzy się średnią wartością współczynników podobieństwa między obiektami należącymi do różnych grup. Oznaczamy przez $S(G_u,G_r)$ jakąkolwiek średnią wartość współczynnika podobieństwa między obiektami należącymi do grupy G_u a obiektami należącymi do grupy G_r.

Postępowanie w metodzie łączonych par jest następujące: zaczyna się od n grup jednoelementowych. Łączy się w grupę parę obiektów najbardziej podobnych. Niech to będą obiekty A_s i A_t, i niech $s < t$. Utworzoną z nich grupę oznaczamy przez $G_s^{(n-1)}$. Następnie w tablicy wskaźników podobieństwa skreśla się wiersze i kolumny odpowiadające obiektom A_s i A_t. Wprowadza się natomiast kolumnę i wiersz dla grupy $G_s^{(n-1)}$, gdzie wpisuje się średnie wskaźniki podobieństwa $S(A_i, G_s^{(n-1)})$ pozostałych obiektów A_i i obiektów należących do grupy $G_s^{(n-1)}$. Tablicę tę oznaczamy przez T_1. Na podstawie tablicy T_1 znajduje się z kolei grupy najbardziej podobne. Mogą to być dwie grupy jednoelementowe lub grupa jednoelementowa i grupa dwuelementowa $G_s^{(n-1)}$. Obie te grupy łączy się w nową

http://rcin.org.pl

grupę. Grupa ta otrzymuje numer równy mniejszemu numerowi łączonych grup i górny wskaźnik równy (n-2). W tablicy T_1 skreśla się wiersze i kolumny odpowiadające obu połączonym grupom i wprowadza wiersz i kolumnę, w które wpisuje się średnie wskaźniki podobieństwa $S(G_i^{(n-1)}, G_u^{(n-1)})$ pozostałych grup do grupy $G_u^{(n-1)}$. Nową tablicę oznaczamy przez T_2.

Przypuśćmy, że w kolejnym kroku mamy m = n-g grup i tablicę średnich wskaźników między nimi T_g. W tablicy tej znajdujemy najbardziej podobne grupy. Niech nimi będą $G_u^{(m)}$ i $G_v^{(m)}$. Grupy te łączymy w nową grupę $G_u^{(m-1)}$. W Tablicy T_g skreślamy wiersze i kolumny odpowiadające grupom $G_u^{(m)}$ i $G_v^{m)}$, wprowadzamy natomiast kolumnę i nowy wiersz dla grupy $G_u^{(m-1)}$, w której wpisuje się obliczone średnie wskaźniki podobieństwa. Postępowanie to powtarza się aż do uzyskania jednej grupy zawierającej wszystkie obiekty.

Przy stosowaniu tej metody można korzystać z różnych sposobów średniowania miar podobieństwa. Najprostszym sposobem jest obliczanie średniej arytmetycznej wskaźników podobieństwa $\bar{s}_{uv}$ między elementami jednej i drugiej grupy zgodnie ze wzorem (24).

Metoda łączonych par może być uogólniona w metodę dopuszczającą jednorazowo łączenie większej ilości obiektów lub grup.

Przykład.

Podziały uzyskane metodą łączonych par Sokala i Michenera podano w tablicy 10. Tablica 11 zawiera pomocnicze obliczenia odległości między tworzonymi grupami. Odległości te obliczane są jako średnie arytmetyczne odległości między obiektami należącymi do obu grup według wzoru (24).

Metoda ta jest łączeniowa. Zaczynamy od podziałów na 12 grup złożonych z poszczególnych powiatów.

Z tablicy odległości znajdujemy parę najpodobniejszych powia-

http://rcin.org.pl

tów. Są nimi powiaty 2 i 3. Oba te powiaty tworzą nową grupę. Oznaczamy ją zgodnie z umową przez $G_2^{(11)}$. W tablicy pomocniczej skreślamy wiersz 2 i 3 a pierwszą kolumnę i pierwszy dodatkowy wiersz oznaczamy przez $G_2^{(11)}$. W kolumnę tę wpisujemy odległości pozostałych powiatów od grupy $G_2^{(11)}$ złożonej z powiatów 2 i 3, obliczonej jako średniej arytmetycznej. Tak więc odległość powiatu 1 od grupy $G_2^{(11)}$ wynosi

$$\frac{3.14 + 3.70}{2} = 3.42.$$

Z tablicy odległości, uzupełnionej kolumną $G_2^{(11)}$ i z pominięciem powiatów 2 i 3 znajdujemy parę najmniej odległych grup. Są to powiat 1 i powiat 4. Oba te powiaty tworzą grupę, której dajemy numer $G_1^{(10)}$. W tablicy dodatkowej skreślamy wiersze odpowiadające powiatom 1 i 4.

Drugą kolumnę i drugi wiersz dodatkowy oznaczamy przez $G_1^{(10)}$. W kolumnę tę wpisujemy odległości do grupy $G_1^{(10)}$. W pierwszym dodatkowym wierszu tej kolumny wpisana jest średnia arytmetyczna odległości powiatów z grupy $G_2^{(11)}$ i z grupy $G_1^{(10)}$, czyli

$$\frac{d(1,2) + d(1,3) + d(2,4) + d(3,4)}{4} =$$

$$\frac{3.141 + 3.703 + 5.151 + 5.688}{4} = 4.42.$$

Najbliższą parą grup jest grupa złożona z powiatów 5 i grupa $G_1^{(10)}$. Powstaje z nich nowa grupa $G_1^{(9)}$. W tablicy pomocniczej skreśla się wiersz odpowiadający powiatowi 5 oraz wiersz odpowiadający grupie $G_1^{(10)}$. W nowej kolumnie, zatytułowanej przez $G_1^{(9)}$, wpisuje się odległości pozostałych grup do tej grupy.

Kontynuowanie postępowania daje dalsze podziały, wyszczególnione w tablicy 12.

http://rcin.org.pl

4.3 Grupowanie przy z góry zadanym kryterium optymalności

Przypuśćmy, że przed grupowaniem mamy sformułowane kryterium optymalności K(P) . Szukamy teraz takiego podziału badanego zbioru obiektów, dla którego funkcja celu K osiąga swoje ekstremum, maksimum lub minimum w zależności od sformułowania kryterium. Znalezienie optymalnego podziału jest w zasadzie możliwe drogą tzw. wyliczania. Mianowicie dokonuje się po kolei wszystkich możliwych podziałów na ustaloną ilość g grup. Dla każdego podziału oblicza się wartość funkcji celu K. Ponieważ ilość możliwych podziałów jest skończona, możliwe jest wskazanie podziału najlepszego w myśl kryterium K. Ze względu jednak na bardzo dużą ilość możliwych podziałów nawet przy umiarkowanie dużej liczbie obiektów, zadanie to wymaga zbyt dużego nakładu pracy rachunkowej lub staje się wręcz niewykonalne nawet przy użyciu bardzo szybkich maszyn liczących.

Sformułowane zadanie upraszcza się, jeżeli ograniczyć się do podziałów hierarchicznych /Ward, 1963/. Zasada postępowania jest następująca. Stosuje się procedurę podziałową lub łączeniową wybierając w każdym kroku ten podział grupy, lub to połączenie dwóch grup, które jest najlepsze w sensie używanego kryterium. Ograniczenie się do metod podziałowych i łączeniowych zmniejsza znacznie ilości możliwych podziałów i upraszcza zadanie. Trzeba jednak zdawać sobie sprawę, że postępowanie to nie musi prowadzić do podziału optymalnego w sensie absolutnym, co więcej, nie wiadomo jak bardzo uzyskany podział różni się od optymalnego.

Inne podejście do zagadnienia grupowania przy z góry zadanym kryterium, to stosowanie różnych metod kolejnych przybliżeń. Ogól-

http://rcin.org.pl

na idea tych metod jest ta, że znajduje się wstępny podział przybliżony na ustaloną ilość grup. Podział ten poprawia się następnie różnymi metodami numerycznymi.

5. PORZĄDKOWANIE LINIOWE

W poprzednich rozdziałach zajmowaliśmy się prawie wyłącznie zagadnieniami klasyfikacji, tj. zagadnieniami związanymi z podziałem zbioru obiektów na grupy jednorodne. W wielu zagadnieniach taksonomicznych powstaje potrzeba uporządkowania badanych obiektów, tzn. przedstawienia w sposób formalny struktury podobieństwa między badanymi obiektami. Każda hierarchia podziałów stanowi takie uporządkowanie. Inne uporządkowanie daje najkrótszy dendryt.

W metodzie Czekanowskiego kolejność obiektów odpowiadająca najlepszemu diagramowi stanowi uporządkowanie liniowe. W uporządkowaniu takim występuje element pierwszy i element ostatni, a dla każdej pary obiektów można orzec, który z nich jest w uporządkowaniu wcześniejszy. Uporządkowanie liniowe jest z wielu względów pożądane. Oryginalna metoda Czekanowskiego jest w praktyce bardzo żmudna i ze względu na brak kryterium wyznaczającego najlepszy diagram nie do zalgorytmowania do obliczeń maszynowych. Powstały ostatnio dwie metody porządkowania liniowego, które są obiektywne i mogą być zaprogramowane na elektroniczne maszyny cyfrowe.

Metoda pierwsza jest pewną formalizacją idei Czekanowskiego /Szczotka, 1972/. W metodzie Czekanowskiego drogą przestawień wierszy i kolumn tablicy miar podobieństwa szuka się takiego diagramu, by wzdłuż głównej przekątnej układały się możliwie duże wskaźniki podobieństwa, a w miarę oddalania się od niej, coraz mniejsze.

http://rcin.org.pl

Te intuicje posłużyły do sformułowania następującego kryterium: miarom podobieństwa przypisuje się wagi w zależności od odległości od głównej przekątnej. I tak, miarom znajdującym się tuż przy przekątnej głównej przypisuje się wagę 1, współczynnikom znajdującym się o dwa kroki od górnej przekątnej wagę 2, współczynnikom oddalonym o trzy kroki od górnej przekątnej wagę 3 i tak dalej. Jako kryterium dobroci uporządkowania uważa się sumę ważoną wszystkich miar podobieństw, czyli w zapisie formalnym

$$Q^{*}(a_1, a_2, \ldots, a_n) = \sum_{v=1}^{p-1} v \sum_{r=1}^{p-v} d_{a_r, a_{r+v}}$$

dla odległości

i $$Q(a_1, a_2, \ldots, a_n) = \sum_{v=1}^{p-1} v \sum_{r=1}^{p-v} s_{a_r, a_{r+v}}$$

dla wskaźników podobieństwa.

Liczby $a_1, a_2, \ldots, a_n$ są numerami obiektów po przestawieniu. Za optymalne uznaje się uporządkowania, dla których Q^{*} jest największe lub Q najmniejsze. Odpowiedni algorytm pozwala znaleźć to uporządkowanie, którego nie można poprawić przez przestawienie dwóch tylko obiektów. Nie musi to być uporządkowanie absolutnie najlepsze, zebrane jednak dotąd doświadczenie wskazuje na to, że uzyskuje się uporządkowania bardzo dobre.

Idea metody drugiej /Z. Piasecki, 1971/ jest następująca. Szuka się dla każdego obiektu A_j takiej liczby y_j, by różnice między wyjściowymi odległościami obiektów d_{ij} a odległościami obliczonymi jako bezwzględne wartości różnic $d'_{ij} = |y_i - y_j|$ były możliwie małe, czyli by

$$F = \sum_{i=1}^{n} \sum_{j=1}^{n} (d'_{ij} - d_{ij})^2 = \text{minimum}.$$

http://rcin.org.pl

Liczby $y_1, y_2, \ldots, y_n$ można interpretować jako współrzędne obiektów na prostej. Ich wielkość porządkuje obiekty liniowo.

6. REDUKCJA ILOŚCI ZMIENNYCH

W niniejszym rozdziale zajmować się będziemy przypadkiem, gdy badane obiekty scharakteryzowane są cechami wielkościowymi.

Zagadnienie klasyfikacji upraszcza się znacznie, jeżeli obiekty scharakteryzowane są jedną tylko cechą, np. X. W tym przypadku wartości cechy porządkują obiekty liniowo. Odległości między obiektami oblicza się jako bezwzględne wartości różnicy wartości cech $d_{ij} = |x_i - x_j|$. Okazuje się, że jeżeli obiekt A_j występuje w uporządkowaniu między obiektami A_i i A_k, to $d_{ik} = d_{ij} + d_{jk}$. Z tej właściwości wynika, że optymalne podziały na grupy jednorodne powstają przez "pocięcie" uporządkowania liniowego. To znaczy, że jeżeli dwa obiekty należą do tej samej grupy, to wszystkie obiekty pośrednie też do niej należą.

W przypadku, gdy obiekty scharakteryzowane są parą cech (x,y), można je przedstawić jako punkty na płaszczyźnie. Otrzymuje się wtedy obraz wzajemnych odległości między obiektami, z którego łatwo może wyniknąć przydatny podział na grupy.

Z przytoczonych argumentów widać korzyści, jakie wynikają z ograniczenia ilości zmiennych opisujących obiekty do jednej lub dwóch, oraz z korzystania z uporządkowania liniowego.

Zajmiemy się najpierw sprawą wykorzystania uporządkowania liniowego do podziału na grupy. W metodzie Czekanowskiego heurystyczny podział na grupy wyłania się w trakcie tworzenia diagramu.

http://rcin.org.pl

Każde z przedstawionych powyżej uporządkowań liniowych można wykorzystać w celu uproszczenia podziału na grupy w przypadku sformułowanego kryterium. Rozważa się mianowicie tylko podziały powstałe przez "pocięcie" uporządkowania i wybiera się ten, który przy ustalonej ilości grup daje optimum w sensie rozważanego kryterium. Ilość wszystkich pocięć, które należy sprawdzić jest znacznie mniejsza niż ilość wszystkich możliwych podziałów i może być sprawdzona w stosunkowo krótkim czasie. Przy takim postępowaniu korzysta się z uporządkowania liniowego przy samym podziale, natomiast wartość kryterium oblicza się korzystając z tablicy miar podobieństwa lub odległości.

Przejdźmy teraz do zagadnienia zastąpienia drogą rachunkową wyjściowego zespołu cech przez mniejszą liczbę nowych zmiennych. Zagadnienie takie nazywa się zagadnieniem redukcji ilości zmiennych lub redukcji wymiaru. Jak już wspomnieliśmy, zainteresowani jesteśmy w redukcji wyjściowego zbioru cech do jednej lub dwóch zmiennych.

Oczywiście redukcję ilości zmiennych chce się przeprowadzić z możliwie małą utratą informacji. Nadają się do tego najlepiej tzw. składowe główne Hotellinga.

Dla przedstawienia ich właściwości będziemy posługiwać się interpretacją geometryczną. Obiekt scharakteryzowany wartościami dwóch cech X_1 i X_2 można przedstawić jako punkt P na płaszczyźnie o współrzędnych (x_1, x_2). Podobnie obiekt scharakteryzowany wartościami trzech cech X_1, X_2 i X_3 będzie reprezentowany przez punkt P w przestrzeni trójwymiarowej o współrzędnych (x_1, x_2, x_3). Formalnie można obiekt scharakteryzowany wartościami P cech X_1, X_2, ..., X_p nazywać punktem P o współrzędnych $(x_1, x_2, \ldots, x_p)$ w przestrzeni o p wymiarach. Mając zbiór punktów przeprowadźmy prostą L i zrzutujmy prostopadle wszystkie punkty P_j na nią. Oznaczmy przez

http://rcin.org.pl

P'_j rzut punktu P_j na prostą L. Długość odcinka (P_j, P'_j) od punktu P_j do jego rzutu P'_j jest równa odległości P_j od prostej L. Za pomocą odpowiednich rachunków można wyznaczyć tę prostą, która jest najbliższa punktom rozważanego zbioru, tzn. taką prostą, że suma kwadratów odległości punktów od niej jest najmniejsza

$$\sum_{j=1}^{n} d^2(P_j, P'_j) = \text{minimum}$$

Na prostej tej wprowadzamy skalę, umieszczając jej początek w środku zbioru i używając takich jednostek jak na pozostałych osiach. Prosta ta nazywa się pierwszą osią główną zbioru punktów. Oznaczmy przez y_{1j} współrzędną punktu P'_j na tej osi. Wartość ta jest określona sumą ważoną wartości (unormowanych cech)

$$y_{1j} = c_{11}x_{1j} + c_{12}x_{2j} + \ldots + c_{1p}x_{pj}$$

Zmienna ta nazywa się pierwszą składową główną zespołu cech $(x_1, \ldots, x_p)$. Ponieważ suma kwadratów odległości między punktami P_j a ich rzutami P'_j jest możliwie mała, tracimy możliwie mało informacji zastępując współrzędne punktów P_j, czyli wartości cech, przez współrzędne punktów P'_j, czyli wartości pierwszej składowej głównej.

Jeżeli posługujemy się wartościami unormowanych cech to średni kwadrat odległości punktów od pierwszej osi głównej jest równy

$$\sum_{i=1}^{p} s_i^2 - \lambda_1 = p - \lambda_1$$

gdzie λ_1 jest równa średniemu kwadratowemu odchyleniu wartości pierwszej składowej głównej, czyli

$$\lambda_1 = \frac{1}{n} \sum_{j=1}^{n} y_{1j}$$

Zarówno wartość λ_1 jak i wektor współczynników

http://rcin.org.pl

$c_1 = (c_{11}, c_{12}, \ldots, c_{1p})$ wyznacza się rachunkowo. Wartość jest tzw. największą wartością własną macierzy korelacji, a wektor c_1 pierwszym wektorem własnym tej macierzy.

Przeprowadźmy przez początek układu prostą L_2 prostopadłą do pierwszej osi głównej L_1. Obie te prostopadłe wyznaczają płaszczyznę. Oznaczmy przez P_j'' rzut punktu P_j, a przez $d(P_j, P_j'')$ odległość punktu P_j od swojego rzutu P_j''. Odległość ta jest jednocześnie odległością punktów P_j od płaszczyzny wyznaczonej przez proste L_1 i L_2. Dobieramy prostą L_2 tak, by suma kwadratów tych odległości była możliwie mała, czyli by

$$\sum_{j=1}^{n} d^2(P_j, P_j'') = \text{minimum}.$$

Prosta L_2 spełniająca ten warunek nazywa się drugą osią główną. Płaszczyzna wyznaczona przez obie osie główne jest spośród wszystkich płaszczyzn płaszczyzną leżącą najbliżej punktów P_j. Jeżeli rozpatrywany zbiór obiektów scharakteryzowany jest dwiema cechami, mamy tylko jedną taką płaszczyznę, mianowicie tę, w której znajduje się układ punktów. Podobnie jak na pierwszej osi głównej wprowadzamy podziałkę na drugiej osi głównej. Współrzędną rzutu punktu P_j na tę oś oznaczymy przez y_{2j}. Jest to tzw. druga składowa główna. Oblicza się ją jako funkcję liniową wartości (unormowanych) cech

$$y_{2j} = c_{21}x_{1j} + c_{22}x_{2j} + \ldots + c_{2p}x_{pj}$$

Wektor współczynników $c_2 = (c_{21}, c_{22}, \ldots, c_{2p})$ jest tzw. drugim wektorem własnym macierzy korelacji cech $X_1, \ldots, X_p$.

Średni kwadrat odległości punktów P_j od płaszczyzny wyznaczonej przez obie pierwsze osie główne wynosi

$$\sum_{j=1}^{n} d^2(P_j, P_j'') = p - \lambda_1 - \lambda_2$$

http://rcin.org.pl

gdzie λ_2 jest średnim kwadratowym odchyleniem drugiej składowej głównej. Jest to druga wartość własna macierzy korelacji. Rzuty punktów rozpatrywanego zbioru na płaszczyznę wyznaczoną przez obie pierwsze osie główne są punktami o współrzędnych równych wartościom obu pierwszych składowych głównych. Ponieważ odległość punktów od tej samej płaszczyzny jest najmniejsza w sensie minimum sumy kwadratów, dlatego też obie składowe główne najlepiej zastępują cały zespół zmiennych.

W wielu przypadkach okazuje się, że położenie punktów w przestrzeni wyznaczonej przez wyjściowy zespół cech unormowanych niewiele odbiega od położenia na płaszczyźnie wyznaczonej przez obie pierwsze składowe główne.

LITERATURA CYTOWANA

Anderberg, M.R., 1958, Cluster Analysis for Application, Wiley, New York.

Baumann, U., 1971, Psychologische Taxonomie, H. Huber, Bern-Stuttgart-Wien.

Bock, H.H., 1974, Automatische Klassifikation, Vandenhoeck u. Rupprecht, Göttingen.

Chojnicki, Z., Czyż, T., 1973, Metody taksonomii numerycznej w regionalizacji geograficznej, PWN, Warszawa.

Cole, A.J. (wyd.), 1969, Numerical Taxonomy, Academic Press, New York.

Czekanowski, J., 1932, "Coefficient of racial likeness" und "durchschnittliche Differenz", Anthrop. Anz., 9, s. 227-249.

Hubert, L., 1973, Monotone invariant clustering procedures, "Psycho-

http://rcin.org.pl

metrika", 38, s. 47-62.

Johnson, S.C., 1967, Hierarchical clustering schemes, "Psychometrika", 32, s. 241-254.

Penrose, L.S., 1954, Distance, size and shape, "Ann. Eugenics", 18, s. 337-343.

Perkal, J., 1953, Taksonomia wrocławska, "Przegląd Antropologiczny, 19, s. 82-96.

Perkal, J., 1960, On the analysing of a set of characteristics, "Zast. Mat.", 5, s. 35-45.

Piasecki, Z., 1971, Nowa metoda taksonomiczna, "Listy Biometryczne", nr 30-33, s. 31-39.

Sokal, R.R., Sneath, P.H., 1963, Principles of Numerical Taxonomy, Freeman Co, San Francisco.

Szczotka, F.A., 1972, On a method of ordering and clustering of objects. "Zast. Mat.", 13, s. 23-33.

Ward, J.H., 1963, Hierarchical grouping to optimize on objective function, "Journ. of the Am. Stat. Assn.", 58, s. 236-244.

http://rcin.org.pl

SPIS TABLIC

strona

Tablica 1 Odległości między powiatami 71

Tablica 2 Hierarchia podziałów uzyskana drogą podziału najkrótszego dendrytu w taksonomii wrocławskiej 72

Tablica 3 Średnie arytmetyczne $\bar{x}$ i odchylenia standardowe s oraz wartości surowe i wartości unormowane 19 cech dla powiatów: augustowskiego /A/, bielskopodlaskiego /BP/ i rybnickiego /R/ 72-73

Tablica 4 Odległości uporządkowane według wielkości . . 73

Tablica 5 Hierarchia podziałów uzyskana metodą najbliższego sąsiada 74

Tablica 6 Hierarchia podziałów otrzymana metodą najdalszego sąsiada 75

Tablica 7 Hierarchia podziałów otrzymana metodą podziałową A 76

Tablica 8 Hierarchia podziałów otrzymana metodą podziałową B 77

Tablica 9 Hierarchia podziałów otrzymana metodą podziałową C 78

Tablica 10 Hierarchia podziałów otrzymana metodą łączonych par 79

Tablica 11 Odległości między grupami dla metody łączonych par 80

http://rcin.org.pl

http://rcin.org.pl

Tablica 1 Odległości między powiatami

Numery x powiatów	1	2	3	4	5	6	7	8	9	10	11	12
1		3.14	3.70	3.11	4.02	5.74	3.78	5.11	5.72	5.65	7.58	9.27
2	3.14		2.40	5.15	5.22	7.90	5.75	5.18	5.24	6.10	7.60	9.66
3	3.70	2.40		5.69	4.96	7.95	6.38	5.14	4.78	5.25	7.52	9.69
4	3.11	5.15	5.69		4.04	4.99	3.81	7.30	6.58	6.05	8.13	8.73
5	4.02	5.22	4.96	4.04		4.24	6.30	6.47	5.07	6.68	9.39	10.39
6	5.74	7.90	7.95	4.99	4.24		6.63	8.97	8.61	8.99	11.48	12.02
7	3.78	5.75	6.38	3.81	6.30	6.63		6.74	7.82	5.65	6.86	8.82
8	5.11	5.18	5.14	7.30	6.47	8.97	6.74		5.08	5.90	7.44	10.48
9	5.72	5.24	4.78	6.58	5.07	8.61	7.82	5.08		5.80	8.36	10.94
10	5.65	6.10	5.25	6.05	6.68	8.99	5.65	5.90	5.80		4.26	7.28
11	7.58	7.60	7.52	8.13	9.39	11.48	6.86	7.44	8.36	4.26		6.12
12	9.27	9.66	9.69	8.73	10.39	12.02	8.82	10.48	10.94	7.28	6.12	

x 1 Starogard
2 Skierniewice
3 Jędrzejów
4 Bydgoszcz
5 Kutno
6 Inowrocław
7 Słupsk
8 Końskie
9 Dębica
10 Sanok
11 Lesko
12 Ustrzyki

http://rcin.org.pl

Tablica 2 Hierarchia podziałów uzyskana drogą podziału najkrótszego dendrytu w taksonomii wrocławskiej

Liczba grup	Podziały
1	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)
2	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11) (12)
3	(1, 2, 3, 4, 5, 6, 7, 8, 9) (10, 11) (12)
4	(1, 2, 3, 4, 5, 6, 7, 9) (8) (10, 11) (12)
5	(1, 2, 3, 4, 5, 6, 7) (8) (9) (10, 11) (12)
6	(1, 2, 3, 4, 5, 6, 7) (8) (9) (10) (11) (12)
7	(1, 2, 3, 4, 5, 7) (6) (8) (9) (10) (11) (12)
8	(1, 2, 3, 4, 7) (5) (6) (8) (9) (10) (11) (12)
9	(1, 2, 3, 4) (5) (6) (7) (8) (9) (10) (11) (12)
10	(1, 4) (2, 3) (5) (6) (7) (8) (9) (10) (11) (12)
11	(1) (2, 3) (4) (5) (6) (7) (8) (9) (10) (11) (12)
12	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12)

http://rcin.org.pl

Tablica 3 Średnie arytmetyczne $\bar{x}$ i odchylenia standardowe s oraz wartości surowe i wartości unormowane 19 cech dla powiatów: augustowskiego /A/, bielskopodlaskiego /BP/ i rybnickiego /R/

Cechy	$\bar{x}$	s	Wartości surowe cech			Wartości unormowane cech		
			A	BP	R	A	BP	R
1. Liczba zatrudnionych na 1 gospodarstwo	1.93	0.51	2.3	2.2	0.5	0.77	0.55	-2.79
2. Powierzchnia użytków rolnych na 1 gospodarstwo	4.77	1.68	8.1	6.3	1,0	1.98	0.91	-2.24
3. Liczba zwierząt w sztukach dużych na 1 gospodarstwo	4.65	1.52	6.7	5.9	0.9	1.36	0.82	-2.47
4. Produkcja globalna na 1 gospodarstwo	89.79	29.03	106.2	91.0	25.4	0.56	0.04	-2.22
5. Liczba zatrudnionych na 100 ha użytków rolnych	48.05	20.63	30.6	36.7	116.7	-0.85	-0.55	3.33
6. Liczba zwierząt pociągowych na 100 ha gruntów ornych	16.14	3.27	16.7	18.2	11.1	0.17	0.63	-1.54
7. Mechaniczna siła pociągowa HP/100 ha gruntów ornych	14.81	8.15	7.1	5.9	27.2	-0.95	-1.09	1.52
8. Nawożenie mineralne kg NPK/1 ha gruntów uprawnych	129.98	32.48	139.8	84.7	167.3	0.30	-1.39	1.15
9. Intensywność i system rolniczego użytkowania ziemi	0.993	0.019	0.99	0.97	1.00	-0.16	-1.19	0.36
10. Procentowy udział upraw trwałych w powierzchni użytków rolnych	1.034	1.395	0.6	0.2	2.2	-0.31	-0.60	0.84
11. Procentowy udział użytków zielonych w użytkach rolnych	21.17	9.33	44.4	28.8	20.9	2.49	0.82	-0.03
12. Intensywność chowu zwierząt sztuk dużych na 100 ha użytków rolnych	75.49	11.37	65.3	70.3	73.1	-0.90	-0.46	-0.21
13. Produktywność ziemi	18 019	3 242	12 202	12 677	22 138	-1.79	-1.65	1.27
14. Produktywność pracy	57 061	16 776	46 472	42 432	53 261	-0.63	-0.87	-0.26
15. Stopień towarowości	44.90	9.03	42.1	43.3	41.2	-0.31	-0.18	-0.41
16. Poziom towarowości	8 253	2 668	5 326	5 496	11 128	-1.10	-1.03	1.08
17. Procentowy udział ludności dwuzawodowej	21.19	11.50	8.1	7.4	59.2	-1.14	-1.20	3.31
18. Procentowy udział produkcji zwierzęcej w produkcji globalnej	44.31	5.06	38.2	48.0	46.4	-1.21	0.73	0.41
19. Procentowy udział produkcji zwierzęcej w produkcji towarowej	68.06	11.22	52.4	78.7	72.7	-1.40	0.95	0.41

http://rcin.org.pl

Tablica 4 Odległości uporządkowane według wielkości

Numery x) powiatów	d	Numery powiatów	d	Numery powiatów	d
2 - 3	2.40	1 - 10	5.64	3 - 11	7.52
1 - 4	3.11	7 - 10	5.65	1 - 11	7.58
1 - 2	3.14	3 - 4	5.69	2 - 11	7.60
1 - 3	3.70	1 - 9	5.72	7 - 9	7.82
1 - 7	3.78	1 - 6	5.74	2 - 6	7.90
4 - 7	3.81	2 - 7	5.75	3 - 6	7.95
1 - 5	4.02	9 - 10	5.80	4 - 11	8.13
4 - 5	4.04	8 - 10	5.90	9 - 11	8.36
5 - 6	4.24	4 - 10	6.05	6 - 9	8.61
10 - 11	4.26	2 - 10	6.10	4 - 12	8.73
3 - 9	4.78	11 - 12	6.12	7 - 12	8.82
3 - 5	4.96	5 - 7	6.30	6 - 8	8.97
4 - 6	4.99	3 - 7	6.38	6 - 10	8.99
5 - 9	5.07	5 - 8	6.47	1 - 12	9.27
8 - 9	5.08	4 - 9	6.58	5 - 11	9.39
1 - 8	5.11	6 - 7	6.63	2 - 12	9.66
3 - 8	5.14	5 - 10	6.68	3 - 12	9.69
2 - 4	5.15	7 - 8	6.74	5 - 12	10.39
2 - 8	5.18	7 - 11	6.86	8 - 12	10.48
2 - 5	5.22	10 - 12	7.28	9 - 12	10.94
2 - 9	5.24	4 - 8	7.30	6 - 11	11.48
3 - 10	5.25	8 - 11	7.44	6 - 12	12.02

x
1 Starogard
2 Skierniewice
3 Jędrzejów
4 Bydgoszcz
5 Kutno
6 Inowrocław
7 Słupsk
8 Końskie
9 Dębica
10 Sanok
11 Lesko
12 Ustrzyki

http://rcin.org.pl

Tablica 5 Hierarchia podziałów uzyskana metodą najbliższego sąsiada

Liczba grup	Podziały											
12	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
11	(1)	(2,	3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
10	(1,	4)	(2,	3)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
9	(1,	2,	3,	4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
8	(1,	2,	3,	4,	7)	(5)	(6)	(8)	(9)	(10)	(11)	(12)
7	(1,	2,	3,	4,	5,	7)	(6)	(8)	(9)	(10)	(11)	(12)
6	(1,	2,	3,	4,	5,	7,	6)	(8)	(9)	(10)	(11)	(12)
5	(1,	2,	3,	4,	5,	6,	7)	(8)	(9)	(10,	11)	(12)
4	(1,	2,	3,	4,	5,	6,	7,	9)	(8)	(10,	11)	(12)
3	(1,	2,	3,	4,	5,	6,	7,	8,	9)	(10,	11)	(12)
2	(1,	2,	3,	4,	5,	6,	7,	8,	9,	10	11)	(12)
1	(1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12)

http://rcin.org.pl

Tablica 6 Hierarchia podziałów otrzymana metodą najdalszego sąsiada

Liczba grup	Podziały											
12	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
11	(1)	(2,	3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
10	(1,	4)	(2,	3)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
9	(1,	4,	7)	(2,	3)	(5)	(6)	(8)	(9)	(10)	(11)	(12)
8	(1,	4,	7)	(2,	3)	(5,	6)	(8)	(9)	(10)	(11)	(12)
7	(1,	4,	7)	(2,	3)	(5,	6)	(8)	(9)	(10,	11)	(12)
6	(1,	4,	7)	(2,	3)	(5,	6)	(8,	9)	(10	11)	(12)
5	(1,	2,	3,	4,	7)	(5,	6)	(8,	9)	(10,	11)	(12)
4	(1,	2,	3,	4,	5,	6,	7)	(8,	9)	(10,	11)	(12)
3	(1,	2,	3,	4,	5,	6,	7)	(8,	9,	10,	11)	(12)
2	(1,	2,	3,	4,	5,	6,	7)	(8,	9,	10,	11,	12)
1	(1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12)

http://rcin.org.pl

Tablica 7 Hierarchia podziałów otrzymana metodą podziałową A

Liczba grup	Podziały											
1	(1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12)
2	(1,	2,	3,	4,	5,	6,	7,	8,	9)	(10,	11,	12)
3	(1,	2,	3,	8,	9)	(4,	5,	6,	7)	(10,	11,	12)
4	(1,	2,	3,	8,	9)	(4,	5,	6,	7)	(10,	11)	(12)
5	(1,	2,	3,	8,	9)	(4,	7)	(5,	6)	(10,	11)	(12)
6	(1,	2,	3)	(8,	9)	(4,	7)	(5,	6)	(10,	11)	(12)
7	(1,	2,	3)	(8)	(9)	(4,	7)	(5,	6)	(10,	11)	(12)
8	(1,	2,	3)	(8)	(9)	(4,	7)	(5,	6)	(10)	(11)	(12)
9	(1,	2,	3)	(8)	(9)	(4,	7)	(5)	(6)	(10)	(11)	(12)
10	(1,	2,	3)	(8)	(9)	(4)	(7)	(5)	(6)	(10)	(11)	(12)
11	(1)	(2,	3)	(8)	(9)	(4)	(7)	(5)	(6)	(10)	(11)	(12)
12	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)

http://rcin.org.pl

Tablica 8 Hierarchia podziałów otrzymana metodą podziałową B

Liczba grup	Podziały
1	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)
2	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11) (12)
3	(1, 2, 3, 4, 5, 6, 7, 8, 9) (10, 11) (12)
4	(1, 2, 3, 4, 5, 6, 7, 9) (8) (10, 11) (12)
5	(1, 2, 3, 4, 5, 6, 7) (8) (9) (10, 11) (12)
6	(1, 2, 3, 4, 5, 7) (6) (8) (9) (10, 11) (12)
7	(1, 2, 3, 4, 5) (6) (7) (8) (9) (10, 11) (12)
8	(1, 4, 5) (2, 3) (6) (7) (8) (9) (10, 11) (12)
9	(1, 4, 5) (2, 3) (6) (7) (8) (9) (10) (11) (12)
10	(1, 4) (2, 3) (5) (6) (7) (8) (9) (10) (11) (12)
11	(1) (2, 3) (4) (5) (6) (7) (8) (9) (10) (11) (12)
12	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12)

http://rcin.org.pl

Tablica 9 Hierarchia podziałów otrzymana metodą podziałową C

Liczba grup	Podziały
1	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)
2	(1, 2, 3, 4, 5, 6, 7, 8, 9) (10, 11, 12)
3	(1, 2, 3, 8, 9) (4, 5, 6, 7) (10, 11, 12)
4	(1, 2, 3, 8, 9) (4, 5, 6, 7) (10, 11) (12)
5	(1, 2, 3, 8, 9) (4, 7) (5, 6) (10, 11) (12)
6	(1, 2, 3) (4, 7) (5, 6) (8, 9) (10, 11) (12)
7	(1, 2, 3) (4, 7) (5, 6) (8) (9) (10, 11) (12)
8	(1, 2, 3) (4, 7) (5, 6) (8) (9) (10) (11) (12)
9	(1, 2, 3) (4, 7) (5) (6) (8) (9) (10) (11) (12)
10	(1, 2, 3) (4) (5) (6) (7) (8) (9) (10) (11) (12)
11	(1) (2, 3) (4) (5) (6) (7) (8) (9) (10) (11) (12)
12	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12)

http://rcin.org.pl

Tablica 10 Hierarchia podziałów otrzymana metodą łączonych par

Liczba grup	Podziały											
12	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
11	(1)	(2,	3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
10	(1,	4)	(2,	3)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
9	(1,	4,	5)	(2,	3)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
8	(1,	4,	5)	(2,	3)	(6)	(7)	(8)	(9)	(10,	11)	(12)
7	(1,	2,	3,	4,	5)	(6)	(7)	(8)	(9)	(10,	11)	(12)
6	(1,	2,	3,	4,	5)	(6)	(7)	(8,	9)	(10,	11)	(12)
5	(1,	2,	3,	4,	5,	8,	9)	(6)	(7)	(10,	11)	(12)
4	(1,	2,	3,	4,	5,	7,	8,	9)	(6)	(10,	11)	(12)
3	(1,	2,	3,	4,	5,	7,	8,	9,	10,	11)	(6)	(12)
2	(1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11)	(12)
1	(1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12)

http://rcin.org.pl

Tablica 11 Odległości między grupami dla metody łączonych par

	$G_2^{(11)}$	$G_1^{(10)}$	$G_1^{(9)}$	$G_{10}^{(8)}$	$G_1^{(7)}$	$G_8^{(6)}$	$G_1^{(5)}$	$G_1^{(4)}$	$G_1^{(3)}$
1	3.42	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	—	—
3	—	—	—	—	—	—	—	—	—
4	5.42	—	—	—	—	—	—	—	—
5	5.09	4.03	—	—	—	—	—	—	—
6	7.92	5.36	4.99	10.23	6.46	8.79	7.12	7.06	7.70
7	6.06	3.80	4.63	6.26	5.35	7.28	5.90	—	—
8	5.16	6.20	6.29	6.67	5.73	—	—	—	—
9	5.01	6.15	5.79	7.08	5.40	—	—	—	—
10	5.67	5.85	6.13	—	—	—	—	—	—
11	7.56	7.85	8.36	—	—	—	—	—	—
12	9.68	9.00	9.46	6.70	9.57	7.90	9.09	9.06	8.58
$G_2^{(11)}$		4.42	4.64	6.62	—	—	—	—	—
$G_1^{(10)}$			—	—	—	—	—	—	—
$G_1^{(9)}$				7.25	—	—	—	—	—
$G_{10}^{(8)}$					6.93	6.87	6.90	6.82	—
$G_1^{(7)}$						5.56	—	—	—
$G_8^{(6)}$									
$G_1^{(5)}$									

http://rcin.org.pl

http://rcin.org.pl

WYKAZ ZESZYTÓW BIULETYNU INFORMACYJNEGO

1. Bibliografia opracowań wykonanych w latach 1970—1973. 1974.
2. Studia nad infrastrukturą społeczną „sensu stricto". 1974.
3. Struktura przestrzenna rolnictwa. Wybór prac. 1975.
4. Wariantowa prognoza rozwoju transportu w Polsce do roku 2000. 1975.
5. Gospodarka wodna. 1975.
6. System regionalnych centrów i ośrodków wzrostu. Materiały z seminarium — Warszawa 22 V 1973. 1975.
7. Wybrane problemy przestrzennego kształtowania infrastruktury społecznej. 1975.
8. Surowik Cz. Wybrane problemy przestrzennego kształtowania infrastruktury ochrony zdrowia i pomocy społecznej. 1975.
9. Prognoza rozwoju infrastruktury ekonomiczno-technicznej Polski do roku 2000. 1975.
10. Uprzemysłowienie i struktura przestrzenna przemysłu. 1976.
11. Nowe metody i techniki w planowaniu przestrzennym. 1976.
12. Problem węzłowy 11.2.1. Podstawy przestrzennego zagospodarowania kraju. Synteza badań I. 1976.
13. Problem węzłowy 11.2.1. Podstawy przestrzennego zagospodarowania kraju. Synteza badań II. 1976.
14. Kształtowanie wewnętrznej struktury jednostek osadniczych. Raport końcowy części B problemu węzłowego 11.2.1. „Zagospodarowanie przestrzenne miast i osiedli wiejskich". 1976.
15. Dziębowski Z. Ekonomiczna efektywność infrastruktury gospodarczej w aglomeracjach miejskich. Raport końcowy tematu. 1976.
16. Bibliografia opracowań wykonanych w latach 1970—1975.