

AKADEMIA EKONOMICZNA IM. OSKARA LANGEGO

WE WROCŁAWIU

WYDZIAŁ ZARZĄDZANIA I INFORMATYKI

MAREK WALESIAK

METODY KLASYFIKACJI W BADANIACH

STRUKTURALNYCH

Praca doktorska napisana

pod kierunkiem

doc. dr hab. Danuty Strahl

WROCŁAW 1985

SPIS TREŚCI

WSTĘP	III
1. TEORETYCZNE PODSTAWY BADAŃ STRUKTURALNYCH	1
1.1. Uwagi wprowadzające	1
1.2. Pojęcie struktury	2
1.3. Klasyfikacja struktur	5
2. KWANTYFIKACJA STRUKTUR	10
2.1. Konstrukcja macierzy danych	10
2.2. Prezentacja wielkości /skali/ oraz kształtu /for- my/ struktur	20
3. MIARY PODOBIENSTWA STRUKTUR	24
3.1. Uwagi wstępne	24
3.2. Przegląd miar podobieństwa struktur	28
3.3. Propozycje miar podobieństwa struktur	32
4. METODY KLASYFIKACJI	40
4.1. Wprowadzenie	40
4.2. Przegląd wybranych metod klasyfikacji	47
4.2.1. Metody hierarchiczne	47
4.2.2. Metody niehierarchiczne	62
4.3. Wyznaczanie optymalnej liczby klas	71
5. REJESTRACJA ZMIAN W CZASIE W WYNIKACH KLASYFIKACJI	77
5.1. Uwagi wstępne	77
5.2. Mierzenie podobieństwa między ciągami klasyfikacji	78

5.3. Mierzenie podobieństwa między podziałami wynikowymi	80
6. ZASTOSOWANIE METOD KLASYFIKACJI DO BADANIA STRUKTURY GAŁĘZIOWEJ PRZEMYSŁU WOJEWÓDZTW POLSKI W ROKU 1975 I 1980	92
6.1. Formalizacja opisu struktur gałęziowych	92
6.2. Klasyfikacja województw Polski ze względu na strukturę gałęziową zatrudnienia w przemyśle	101
6.2.1. Podobieństwo wielkości /skali/	101
6.2.2. Podobieństwo kształtu /formy/	115
6.2.3. Rejestracja zmian typologicznych w strukturze gałęziowej zatrudnienia w przemyśle województw Polski	128
6.3. Klasyfikacja województw Polski ze względu na strukturę gałęziową podstawowych relacji ekonomicznych w przemyśle	130
6.3.1. Podobieństwo wielkości /skali/	130
6.3.2. Podobieństwo kształtu /formy/	144
6.3.3. Rejestracja zmian typologicznych w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle województw Polski	156
7. WNIOSKI KOŃCOWE	158
7.1. Wnioski natury metodycznej	158
7.2. Wnioski wynikające z badań empirycznych	161
LITERATURA	163

WSTEP

Z powodu wyczerpywania się ilościowych czynników rozwoju / w okresie ich dominacji następują stosunkowo małe zmiany strukturalne/ koniecznością stają się istotne przemiany strukturalne w gospodarce narodowej, a szczególnie w przemyśle ze względu na jego decydujący wpływ na poziom i tempo rozwoju gospodarczego. Dzięki tym przemianom można bowiem wyzwolić dodatkowe rezerwy tkwiące w danej gospodarce. Aby tego dokonać konieczne są badania mające na celu wykrycie mechanizmu przekształceń strukturalnych. Dla realizacji tych zadań niezbędnym wydaje się być wypracowanie odpowiedniego podejścia metodycznego.

Jednym z narzędzi badań strukturalnych są metody klasyfikacji. Pozwalają one na badanie przestrzennego i dynamicznego zróżnicowania obiektów gospodarczych ze względu na ich strukturę.

Niniejsze opracowanie poświęcono właśnie zastosowaniu metod klasyfikacji do badań strukturalnych. Praca ma przede wszystkim charakter metodyczny. Występuje w niej również część empiryczna celem potwierdzenia użyteczności praktycznej przyjętej metodyki badawczej. Specyfika badań strukturalnych wymaga szczególnego potraktowania takich zagadnień związanych z metodami klasyfikacji, jak sporządzanie macierzy danych, normowanie cech strukturalnych, ocena podobieństwa struktur. Inne zagadnienia, jak metody klasyfikacji oraz rejestracja zmian w czasie w wynikach klasyfikacji mają charakter uniwersalny, tzn. nie wynikają ze specyfiki podejścia strukturalnego, a zostały jedynie zaadaptowane dla potrzeb badań strukturalnych.

często w sferze ekonomii występuje sytuacja, w której badana populacja jest skończona, a badaniu podlegają wszystkie obiekty. Nie występuje wtedy problem powtórzenia eksperymentu. W takim układzie autor zdecydował się zastosować do zagadnienia klasyfikacji podejście deterministyczne, w którym zakłada się, że dany jest skończony zbiór n obiektów na których dokonuje się obserwacji m zmiennych nielosowych nie obarczonych błędami pomiaru.

Praca składa się z dwóch zasadniczych części: teoretycznej /rozdziały 1 - 5/ oraz empirycznej /rozdział 6/.

Rozdział pierwszy pracy zawiera teoretyczne podstawy badań strukturalnych. Podana została krótka historia badań strukturalnych w kraju i zagranicą. Następnie autor przeprowadził analizę funkcjonujących definicji struktury i na tym tle sformułował własną koncepcję struktury. Rozdział kończą rozważania dotyczące klasyfikacji struktur. Zaprezentowano przegląd spotykanych w literaturze kryteriów klasyfikacji struktur, jak i własne propozycje w tym zakresie.

W rozdziale drugim przedstawiono konstrukcję macierzy danych, która to stanowi ważny element badań strukturalnych. Zdefiniowano takie podstawowe pojęcia, jak obiekt, struktura obiektu, cecha strukturalna, elementy struktury oraz pojęcia od nich pochodne. Dużo miejsca poświęcono również omówieniu skal pomiaru cech strukturalnych. Zasadnicza część tego rozdziału traktuje o prezentacji wielkości /skali/ oraz kształtu /formy/ struktur. Autor przedstawia w nim własne propozycje normowania cech strukturalnych.

W rozdziale trzecim zaprezentowano miary podobieństwa struktur. Wprowadzenie do tego rozdziału zostało poświęcone omówieniu podziału tych miar na miary odległości i bliskości oraz przedstawieniu warunków, które winny spełniać miary podobieństwa. Autor wprowadza również dodatkowy warunek dla miar podobieństwa struktur obiektów. Warunek ten znacznie ogranicza liczbę możliwych do za-

stosowania miar w badaniach strukturalnych. Ponadto z przeglądu miar podobieństwa wynika, że niektóre z nich mogą być zastosowane tylko przy badaniu podobieństwa kształtu struktur. Są wśród nich również miary posiadające pewne wady. Wszystko to skłoniło autora do zaprezentowania własnych propozycji miar odległości oraz bliskości struktur.

Rozdział czwarty zawiera omówienie wybranych metod klasyfikacji. W pracy zastosowano ogólny podział metod klasyfikacji, mianowicie podział na metody hierarchiczne i niehierarchiczne oraz na metody aglomeracyjne i deglomeracyjne. Taki podział będzie w pełni wystarczający, jako że w niniejszej pracy autor ogranicza się do przeglądu metod klasyfikacji opartych wyłącznie na macierzy podobieństw. W niektórych algorytmach metod klasyfikacji przeprowadzono pewne modyfikacje celem uczynienia ich bardziej zrozumiałymi. Ponadto przedstawiono takie pojęcia, jak klasyfikacja i klasa homogeniczna oraz omówiono proponowane w literaturze wybrane własności formalne i praktyczne metod klasyfikacji. Końcowa część tego rozdziału zawiera omówienie sposobów wyznaczania optymalnej liczby klas, na które należy podzielić zbiór badanych obiektów. Problem ten jest ważny, gdyż użytkownika metod klasyfikacji interesują przede wszystkim wyniki klasyfikacji, a nie sama zastosowana metoda.

W rozdziale piątym przedstawiono sposoby rejestracji zmian w czasie w wynikach klasyfikacji obiektów. Omówiono krótko sposoby mierzenia podobieństwa między ciągami klasyfikacji. Główne rozważania zawarto w podrozdziale 5.3, który jest poświęcony sposobom mierzenia podobieństwa między podziałami wynikowymi. Autor zaproponował warunki jakie powinna spełniać dobrze określona miara i na tym tle dokonał analizy różnych miar, proponując jedną z nich jako najlepszą z punktu widzenia tych własności.

Rozdział szósty obejmuje badania empiryczne, które stanowią

ilustrację części metodycznej. Metody klasyfikacji zastosowano do badania struktur gałęziowych przemysłu województw Polski. Dokonano sformalizowanego opisu struktur gałęziowych. W badaniu jako obiekty przyjęto województwa Polski wyodrębnione jako jednostki administracyjne 1.06. 1975 r. Jako elementy struktury przyjęto grupy gałęzi przemysłu. Badanie polegało na klasyfikacji województw Polski ze względu na gałęziowy poziom oraz profil uprzemysłowienia. Przeprowadzenie tych badań wymagało doboru odpowiedniego zestawu cech w klasyfikacji wielocechowej oraz ustalenia priorytetu cechy w klasyfikacji jednocechowej. W pracy przy doborze cech wykorzystano podejście merytoryczne. Na zakończenie części wstępnej tego rozdziału przedstawiono ogólny schemat postępowania metodycznego przy badaniu struktur gałęziowych przemysłu województw Polski w roku 1975 i 1980. Następnie przedstawiono i zinterpretowano otrzymane klasyfikacje województw Polski ze względu na gałęziowy poziom oraz profil uprzemysłowienia dla lat 1975 i 1980. Wyniki badań uzupełniono ich analizą oraz graficzną prezentacją.

W ostatnim rozdziale pracy, stanowiącym jej podsumowanie, sformułowano wnioski natury metodycznej oraz wnioski wynikające z badań empirycznych. Wnioski natury metodycznej wskazują z jednej strony na praktyczne korzyści wynikające z zastosowania do badań strukturalnych metod klasyfikacji, z drugiej strony zaś wskazują na konieczność ciągłego doskonalenia technik statystyczno-ekonometrycznych związanych z zagadnieniem klasyfikacji. Z uwagi na to, że wiele decyzji związanych z metodami klasyfikacji ma charakter subiektywny, od badacza problemu wymaga się wiedzy w zakresie wybranej dyscypliny badawczej oraz w zakresie metod statystyczno-ekonometrycznych.

1. TEORETYCZNE PODSTAWY BADAŃ STRUKTURALNYCH

1.1. Uwagi wprowadzające.

Złożoność zagadnienia omawianego w niniejszym rozdziale najlepiej wyrazi cytat z pracy P.S. Florence'a [28, s.22] : "Struktura to słowo, które się łatwo pojmuje, ale i łatwo zatracą się jego znaczenie; niemal stale słyszymy i czytamy o strukturze płac, strukturze kapitału, strukturze klasowej, strukturze wieku, tak jak o strukturze anatomicznej, a nawet o strukturze powieści. Mimo swej pozornej różnorodności wyrażenia te mają istotną wspólną treść".

Badania strukturalne na większą skalę rozpoczęły się w pierwszej połowie XX wieku, kiedy pojęcie to wprowadził do teorii ekonomii politycznej E. Wagemann w pracy [105]¹. Kulminacyjny okres przypadł na lata pięćdziesiąte i sześćdziesiąte, a znalazło to wyraz w wielu pracach zarówno teoretycznych jak i empirycznych, głównie jednak u ekonomistów zachodnich [3; 26; 53; 57; 73; 74]. W Polsce do bardziej znanych pozycji w zakresie badań strukturalnych należą prace A. Karpińskiego [46], J. Lisikiewicza [54; 97], S. Marciniaka [58; 59], B. Minca [63; 64], J. Tobera [103] oraz L. Zachera [119].

Na wielość definicji struktury - różniących się między sobą - wskazują m. in.: M. Strihafka w pracy [95], przedstawiając chronologiczny rozwój pojęcia struktury, Z. Szymła w pracy [101], ujmując w skrócie rozmaite definicje struktury, oraz autorka pracy [48] stwierdzając, że "uderza nie tyle ilość badaczy zajmujących się tym zagadnieniem, ile różnice w definicjach i to w zakresie tej samej dyscypliny".

¹Por. [95, s.55].

Na początku rozważań należy podkreślić, że termin "struktura" ma charakter interdyscyplinarny, nie ma więc potrzeby odmiennego ujmowania struktury przez różne dziedziny badawcze. W stosunku do pojęcia struktury np. struktura gospodarcza jest czymś wtórnym. Nie wydaje się celowe jej definiowanie, ponieważ nie ma różnicy między ogólnym pojęciem struktury a pojęciem struktury gospodarczej. Przymiotnik "gospodarcza" wskazuje, jakiej sfery działalności ludzkiej dana struktura dotyczy.

1.2. Pojęcie struktury.

Na wysokim poziomie abstrakcji posługujemy się pojęciami, którym często nie odpowiada żaden istniejący obiekt. Dopiero w procesie poznania, uchylając kolejne założenia, konkretyzując pojęcia abstrakcyjne, otrzymuje się sytuacje bardziej realistyczne. Prowadzi to zwykle do rezygnowania z pełnej informacji o interesujących nas zjawiskach. Mamy wówczas do spełnienia trzy zadania. Pierwsze z nich odpowiada tworzeniu pojęcia struktury, drugie sprowadza się do odpowiedzi na pytanie czy konkretna sytuacja /obiekt badania/ może być przyporządkowana ustalonemu pojęciu, trzecie - po uzyskaniu twierdzącej odpowiedzi na pytanie drugie - polega na przedstawieniu obrazu badanej struktury obiektu w formie liczbowej.

W wyniku tych trzech kroków otrzymuje się liczbowy obraz struktury obiektu, która może być w tej formie poddana badaniu statystyczno-ekonomicznemu.

Realizacja pierwszego zadania polega na próbie stworzenia pojęcia struktury. Na podstawie definicji różnych autorów, uznawanych powszechnie za reprezentatywne, określone zostaną warunki uzasadniające nazywanie jakiejś konstrukcji strukturą [110].

Pierwszy warunek, co do którego w zasadzie wszyscy autorzy definicji są zgodni, przyjmuje następujące brzmienie : struktura jest

to całość, która składa się z elementów. Aspekt ten uwzględniają w swoich definicjach m. in. A. Marchal [57, s.77]², według którego "struktura to elementy pewnej całości", J. Piaget [74, s.34] - "struktura ma /.../ charakter całości /.../, struktura niewątpliwie jest utworzona z elementów", O. Kryn i P. Pelikan [52, s.21] - "strukturę układu tworzą elementy z których składa się układ", J. Tober [103, s.24] - "struktura /.../ jest /.../ określoną całością zbudowaną /.../ z określonych elementów", J. Lisikiewicz [54, s.16] - "struktura obejmuje zespół elementów tworzących całość".

W tym miejscu należy również podkreślić, że całość w strukturze składa się z co najmniej dwóch elementów struktury, a zbiór tychże elementów spełnia warunki rozłączności /elementy struktury nie mają części wspólnych/ oraz zupełności /suma elementów struktury tworzy całość/.

Drugi warunek sprowadza się do tego, że struktura obrazuje układ zależności między elementami badanej całości oraz między poszczególnymi elementami a całością. Zależności te zwane prawami składania, nadają całości własności odrębne od własności elementów. Takie ujęcie - ogólnie mówiąc - reprezentują J. Akerman [3]³, który stwierdza, że "pojęcie struktury odnosi się do związku między częściami danej całości a także związków między tymi częściami a całością", J. Piaget [74, ss. 34,36] pisząc, że "struktura niewątpliwie jest utworzona z elementów, ale te są podporządkowane prawom cechującym system jako taki; owe prawa, zwane prawami składania, nie dają się sprowadzić do łączących kojarzeń, lecz nadają całości jako takiej, własności zbioru odrębne od własności elementów" ; "całość jest czymś odmiennym od zwykłej sumy poprzedzających ją elementów", J. Lisikiewicz [54, ss. 16-17] - według którego "struktura obejmuje /.../ relacje zachodzące między elementami a całością" czy też inni przedstawiciele, jak O. Kryn i P. Pelikan [52],

²Cyt. za M. Strihańką [95, s.59] .

³Tamże s.56.

A. Karpiński [46] oraz J. Tober [103] .

Większość autorów w swoich definicjach kładzie wyraźny akcent na zależności między elementami badanej całości oraz między poszczególnymi elementami a całością. Pomija się zatem rolę /funkcje/ jakie odgrywają w strukturze jej poszczególne elementy. Pojawia się w związku z tym pytanie czy struktury są systemami przekształceń /dynamiczne/ czy jakąś formą statyczną. Jak słusznie stwierdza J. Piaget [74, s.38] : "wszystkie znane struktury /.../ są faktycznie systemami przekształceń, /.../, gdyby nie zawierały takich przekształceń, upodobniałyby się do czysto statycznej formy i straciły całą wartość wyjaśniającą". Struktura jest więc systemem przejściowym, zmieniającym się w czasie, a proces zachodzących w niej przemian tworzy całość dzięki pewnemu związkowi zależności, który sprawia, że "nowa" struktura nie może powstać w oderwaniu od "starej". Właśnie ruch w czasie poszczególnych elementów całości zachowujących względną samodzielność prowadzi do zmian w "istniejących" strukturach i do powstania "nowych" struktur.

Należy jednak pamiętać, jak słusznie zauważa B. Minc w pracy [63, s.93] , że: "Każda struktura jest makrostrukturą wobec swoich części i mikrostrukturą w stosunku do struktury, której część stanowi. Mikrostruktury posiadają względną samodzielność wobec makrostruktur, są im jednak w jakiś sposób podporządkowane".

Drugie zadanie wymagało odpowiedzi na pytanie, czy konkretna sytuacja może być przyporządkowana ustalonemu pojęciu struktury. W rozważaniach nad tym problemem można posłużyć się przykładem. Najpierw należy wybrać obiekt badania a następnie na nim określić strukturę. Określenie struktury na obiekcie wymaga odpowiedzi na pytania: a/ jakiej dziedziny badana struktura będzie dotyczyć, b/ co rozumie się przez całość w danej dziedzinie, c/ jakie elementy tej całości należy wyodrębnić. Niech przedmiotem badania będzie struktura gospodarcza województwa, w której całością jest przemysł, będący działem gospodarki narodowej, elementami zaś - grupy gałęzi

przemysłu. Mając tak określoną strukturę obiektu należy wybrać problem badawczy z jej zakresu oraz ustalić zespół cech /cechę/, który pozwoli rozstrzygnąć ten problem. Zakładając, że tym problemem będzie poziom uprzemysłowienia województwa można przyjąć jako cechę jeden z następujących mierników [122, s.15] :

- wielkość produkcji,
- liczba zatrudnionych,
- wartość środków trwałych.

Zadanie trzecie polega na przedstawieniu liczbowego obrazu struktury obiektu. Wymaga ono usytuowania tejże struktury obiektu w czasie i w przestrzeni. Niech będzie to struktura gałęziowa przemysłu w województwie jeleniogórskim w roku 1980. Ostatnim etapem jest zebranie danych statystycznych z materiałów źródłowych.

Ramowy schemat opisanego postępowania przedstawia rys. 1.1.

Oczywiście prowadzenie wszelkich badań porównawczych wymaga określenia nie jednej struktury obiektu ale zbioru struktur obiektów /patrz rozważania zawarte w rozdziale drugim/.

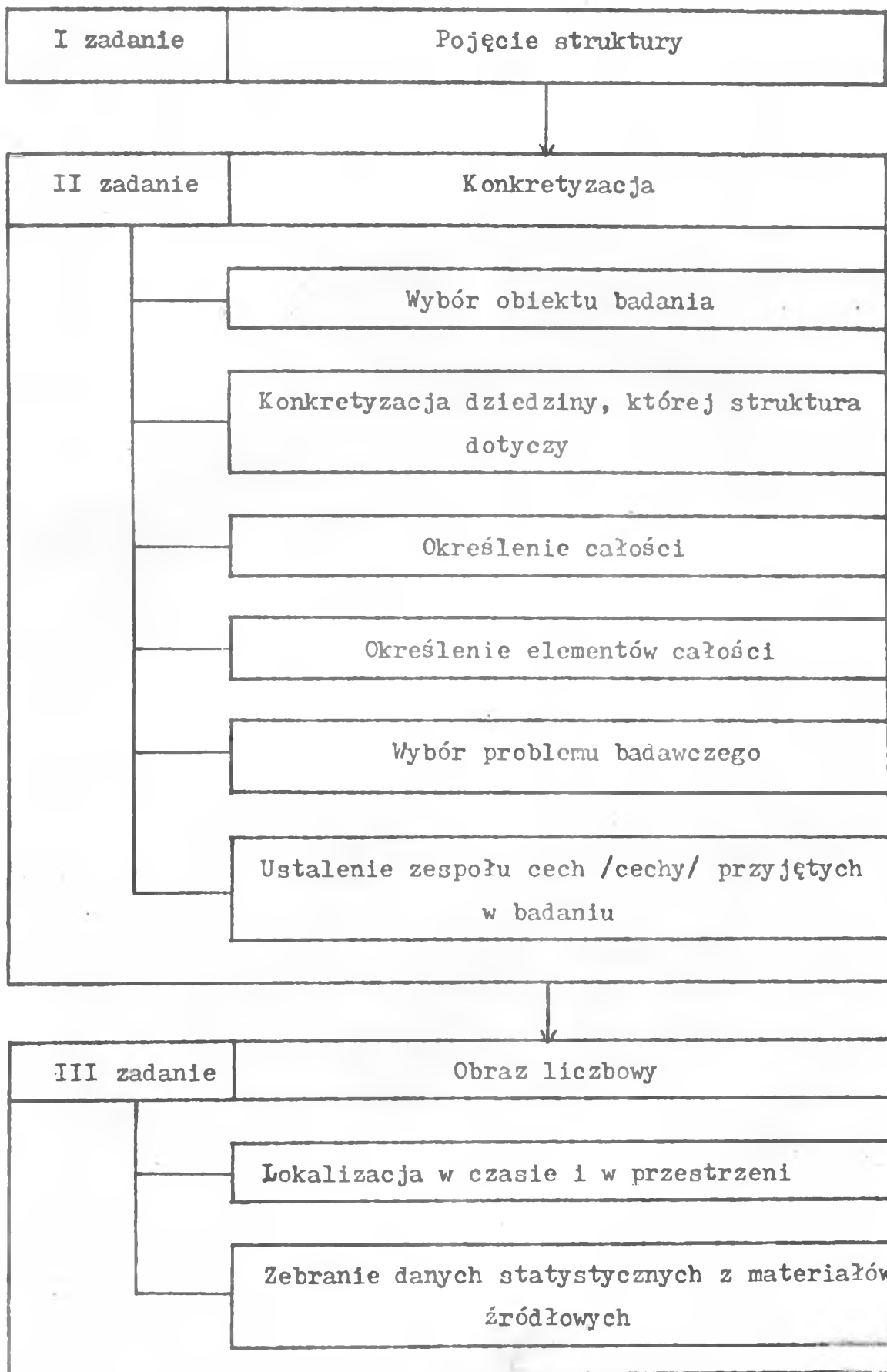
1.3. Klasyfikacja struktur.

Omawiając pojęcie struktury stwierdzono, że stosuje się je we wszystkich dziedzinach badań, a więc że ma ono charakter interdyscyplinarny. Z tego względu struktury można sklasyfikować zależnie od rodzaju nauk, w których one występują, na : matematyczne, fizyczne, chemiczne, psychologiczne, socjologiczne, ekonomiczne /gospodarcze/ itd.

Z punktu widzenia stopnia agregacji i dezagregacji badanych struktur wyodrębnia się [55] :

- a/ makrostruktury,
- b/ mikrostruktury.

/I/



Rys. 1.1. Schemat postępowania w badaniach strukturalnych

Makrostruktury służą do określania relacji zachodzących między elementami struktury opisującymi obiekty wyższych rzędów, w przeciwieństwie do mikrostruktur, które służą opisowi związków występujących między wielkościami odnoszącymi się do obiektów niższych rzędów. Na przykład w przemyśle układami makrostrukturalnymi są relacje między poszczególnymi gałęziami przemysłu, a układami mikrostrukturalnymi w stosunku do poprzednich - relacje między rodzajami przemysłu czy grupami wyrobów w ramach poszczególnych gałęzi.

W związku z tym, że elementy struktury mogą się różnić rodzajem, czasem i miejscem swego występowania, struktury można podzielić na [56, s.805] :

- a/ rodzajowe /rzeczowe/,
- b/ czasowe, /II/
- c/ przestrzenne /terytorialne/.

Struktury rodzajowe opisują relacje zachodzące w danym czasie i miejscu między elementami struktury wyróżniającymi się swoim rodzajem.

Struktury czasowe opisują stosunki między elementami struktury różniącymi się czasem swego występowania, a jednakowymi pod względem swego rodzaju i lokalizacji.

Struktury przestrzenne opisują relacje między elementami struktury różniącymi się miejscem swego występowania, a takimi samymi pod względem rodzaju i czasu.

Wprowadzenie obrazu⁴ struktury danego obiektu pozwala sklasyfikować je z dwóch punktów widzenia. Przyjmując za punkt widzenia stopień złożoności badanego zjawiska wśród struktur wyodrębnia się [110] :

- a/ proste, /III/
- b/ złożone.

Podstawowa różnica między nimi sprowadza się do tego, że struktury proste opisane są za pomocą jednej cechy, natomiast struktury zło-

żone za pomocą zespołu cech.

Z kolei sposób prezentacji danych liczbowych pozwala podzielić struktury na:

a/ prezentujące wielkość /skalę/,

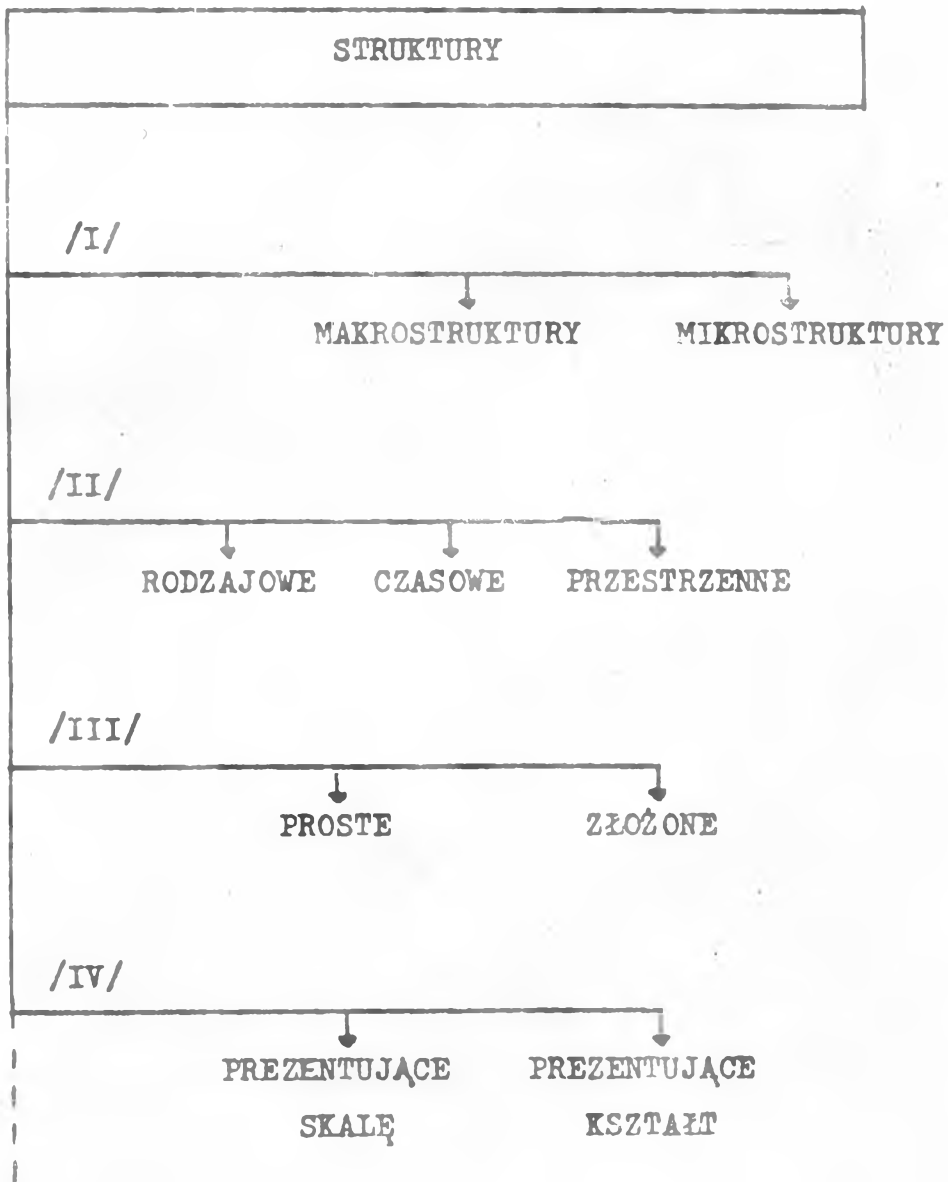
/IV/

b/ prezentujące kształt /formę/.

Struktura prezentuje skalę, gdy po unormowaniu cech strukturalnych ich składowe wyrażają poziom danej cechy. Z kolei prezentuje kształt, wtedy gdy po unormowaniu cech strukturalnych ich składowe wyrażają proporcje między poszczególnymi składowymi danej cechy. Sposoby normowania cech strukturalnych umożliwiające prezentację wielkości /skali/ oraz kształtu /formy/ struktur przedstawione zostaną w podrozdziale 2.2.

Przedstawione klasyfikacje struktur nie wyczerpują z pewnością zbioru wszystkich możliwych podziałów. Stanowią jednak wystarczającą i dogodną podstawę do dalszych rozważań zawartych w niniejszej pracy. Klasyfikacje oznaczone numerami I - IV w sposób syntetyczny przedstawia rys. 1.2.

⁴ Patrz rozważania zawarte w podrozdziale 2.1.



Rys. 1.2. Schemat klasyfikacji struktur

2. KWANTYFIKACJA STRUKTUR

2.1. Konstrukcja macierzy danych.

Punktem wyjścia w badaniach strukturalnych jest tzw. macierz danych. Konstrukcja tej macierzy rozpoczyna się od określenia trzech zbiorów: A , M i K . Pierwszy z nich o elementach A_r dla $r = 1, 2, \dots, n$ nazywa się zbiorem obiektów badania, drugi o elementach M_l dla $l = 1, 2, \dots, m$ - zbiorem cech strukturalnych, trzeci zaś o elementach K_j dla $j = 1, 2, \dots, k$ nazywa się zbiorem elementów struktury.

Z punktu widzenia badań strukturalnych, podobnie jak i każdego innych badań, zasadniczym zagadnieniem jest określenie elementarnej jednostki badawczej czyli obiektu badania. W pracy przez obiekt rozumie się "najmniejszy element poddany obserwacji, który dostarcza podstawowej z punktu widzenia sformułowanej hipotezy informacji" [91, ss.19-20]. Obiekty rozumiane są zarówno w sensie dosłownym, jak i przenośnym: Obiektem jest więc określona rzecz, osoba, wytwór myślowy, zjawisko lub zdarzenie. Konkretnymi przykładami obiektów są: województwo jeleniogórskie, miasto Jelenia Góra, konkretna idea, konkretny człowiek itd.

W każdym obiekcie badania $A_r \in A$ należy określić strukturę składającą się z k elementów. Strukturę r -tego obiektu $S/A_r/$ przedstawia wektor $[K_j/A_r/]$:

$$[K_j/A_r/] = \begin{bmatrix} K_1/A_r/ \\ K_2/A_r/ \\ \vdots \\ K_k/A_r/ \end{bmatrix}$$

/2.1/

gdzie: $K_j/A_r/$ - j-ty element struktury w r-tym obiekcie.

W celu lepszego zobrazowania istoty problemu można posłużyć się przykładem. Niech A_r oznacza województwo jeleniogórskie a zbiór K grupy gałęzi przemysłu. Stąd

$$\left[K_j/A_r/ \right] = \begin{bmatrix} K_1/A_r/ \\ K_2/A_r/ \\ \vdots \\ K_9/A_r/ \end{bmatrix} = \quad /2.2/$$

przemysł paliwowo-energetyczny w woj. jeleniogórskim przemysł metalurgiczny w woj. jeleniogórskim ----- pozostałe gałęzie przemysłu w woj. jeleniogórskim
--

Przyporządkowanie każdemu obiektowi $A_r \in A$ struktury składającej się z k elementów zapisuje się jako następujące odwzorowanie

$$F : A \text{ ----> } S/A/ \quad \text{zbiory} \quad /2.3/$$

$$A_r \text{ ----> } S/A_r/ = \left[K_j/A_r/ \right]_{k \times 1} \quad \text{elementy}$$

gdzie: $S/A/$ - zbiór struktur obiektów; $F/A_r/ = S/A_r/ = \left[K_j/A_r/ \right]$.

W związku z tym, że bezpośrednio porównywanie struktur obiektów jest niemożliwe - ze względu na złożoność ich natury - wprowadza się pojęcie obrazów struktur obiektów, które w sposób pośredni umożliwiają ich porównanie. Określenie obrazu struktury obiektu jest możliwe przez wprowadzenie cechy strukturalnej¹, którą interpretuje się w niniejszej pracy jako funkcję odwzorowującą zbiór struktur obiektów w zbiór ich obrazów. W ujęciu formalnym cecha to

¹Zamiennie będzie używany termin cecha.

odwzorowanie

$$I_1 : S/A/ \dashrightarrow E_1^k \subset R^k \quad \text{zbiory}$$

/2.4/

$$S/A/_r/ \dashrightarrow x_{.lr}^T \quad \text{elementy}$$

$$\text{gdzie: } I_1 [S/A/_r/] = x_{.lr}^T = (x_{1lr} \ x_{2lr} \ \dots \ x_{klr})^T \in E_1^k ;$$

E_1^k - zbiór wektorów realizacji /obrazów/ o wymiarach $k \times 1$;

T - znak transpozycji.

W taki oto sposób otrzymuje się ostatecznie blokową macierz danych X_{\dots} / $k \times n \times m$ /:

x_{111}	x_{121}	\dots	x_{1m1}
x_{211}	x_{221}	\dots	x_{2m1}

x_{k11}	x_{k21}	\dots	x_{km1}

x_{112}	x_{122}	\dots	x_{1m2}
x_{212}	x_{222}	\dots	x_{2m2}

x_{k12}	x_{k22}	\dots	x_{km2}

\vdots	\vdots		\vdots

x_{11n}	x_{12n}	\dots	x_{1mn}
x_{21n}	x_{22n}	\dots	x_{2mn}

x_{k1n}	x_{k2n}	\dots	x_{kmn}

$$= \begin{bmatrix} X_{..1} \\ X_{..2} \\ \vdots \\ X_{..n} \end{bmatrix} = X_{\dots}$$

/2.5/

gdzie: $i, j = 1, 2, \dots, k$ /numer elementu struktury/,

$r, s = 1, 2, \dots, n$ /numer obiektu/,

$l = 1, 2, \dots, m$ /numer cechy strukturalnej/,

x_{ilr} - wartość liczbowa i -tego elementu struktury zmierzona dla l -tej cechy w r -tym obiekcie,

$\mathbb{X}_{..r}$ - liczbowy obraz struktury obiektu A_r ,

$\mathbb{X}_{.lr}^T = (x_{1lr} \ x_{2lr} \ \dots \ x_{klr})^T$ - wektor obserwacji na l -tej cesze strukturalnej w strukturze r -tego obiektu,

$\mathbb{X}_{1.r} = (x_{11r} \ x_{12r} \ \dots \ x_{1mr})$ - wektor obserwacji na i -tym elemencie w strukturze r -tego obiektu.

Macierze $\mathbb{X}_{..r}$ można uważać za reprezentacje lub obrazy struktur obiektów $S/A_r/$. Należy to rozumieć jako odwzorowanie

$$I : S/A/ \dashrightarrow G_1^{mk} \subset R^{mk} \quad \text{zbiory} \quad /2.6/$$

$$S/A_r/ \dashrightarrow \mathbb{X}_{..r} \quad \text{elementy}$$

$$\text{gdzie: } I [S/A_r/] = \mathbb{X}_{..r} = (\mathbb{X}_{.1r}^T \ \mathbb{X}_{.2r}^T \ \dots \ \mathbb{X}_{.mr}^T) \in G_1^{mk};$$

G_1^{mk} - przestrzeń obrazów struktur obiektów.

Odwzorowanie zbioru obiektów A w przestrzeń obrazów struktur obiektów G_1^{mk} zapisuje się następująco

$$L : A \dashrightarrow G_1^{mk} \subset R^{mk} \quad \text{zbiory} \quad /2.7/$$

$$A_r \dashrightarrow \mathbb{X}_{..r} \quad \text{elementy}$$

$$\text{gdzie: } L/A_r/ = \mathbb{X}_{..r}; \quad L \text{ jest superpozycją funkcji } I \circ F.$$

Ze względu na wygodę i pewne uzasadnienie logiczne obraz struktury obiektu często utożsamia się ze strukturą obiektu lub z sa-

mym obiektem.

Kolumny macierzy blokowej \mathbb{X}_{\dots} mogą być interpretowane jako reprezentacje lub obrazy cech. Należy to rozumieć jako odwzorowanie

$$S : M \longrightarrow G_2^{nk} \subset R^{nk} \quad \text{zbiory} \\ M_1 \longrightarrow \mathbb{X}_{\cdot 1} \quad \text{elementy} \quad /2.8/$$

$$\text{gdzie: } S/M_1/ = \mathbb{X}_{\cdot 1} = (\mathbb{X}_{\cdot 11} \mid \mathbb{X}_{\cdot 12} \mid \dots \mid \mathbb{X}_{\cdot 1n})^T \in G_2^{nk};$$

G_2^{nk} - przestrzeń obrazów cech strukturalnych.

Wiersze macierzy blokowej \mathbb{X}_{\dots} mogą być interpretowane jako reprezentacje lub obrazy elementów struktury. Należy to rozumieć jako odwzorowanie

$$T : K \longrightarrow G_3^{mn} \subset R^{mn} \quad \text{zbiory} \\ K_j \longrightarrow \mathbb{X}_{1..} \quad \text{elementy} \quad /2.9/$$

$$\text{gdzie: } T/K_j/ = \mathbb{X}_{1..} = (\mathbb{X}_{1.1} \mid \mathbb{X}_{1.2} \mid \dots \mid \mathbb{X}_{1.n}) \in G_3^{mn};$$

G_3^{mn} - przestrzeń obrazów elementów struktury².

W analizie porównawczej występują dwa rodzaje badań, tzn. nad obiektami oraz nad cechami [77]. Z tego względu celowe wydaje się być podanie interpretacji geometrycznej obrazu struktury obiektu i obrazu cechy. W przypadku, gdy zbiór cech strukturalnych jest jednoelementowy / $m = 1$ / to obrazem struktury obiektu w interpretacji geometrycznej jest wektor lub punkt w przestrzeni k -wymiarowej. Jeśli zbiór cech jest m -elementowy / $m \geq 2$ / to obrazem struktury obiektu jest wektor lub punkt w przestrzeni / mk -wymiarowej.

²Por. rozważania zawarte w pracach [39; 77].

Z kolei obrazem każdej cechy strukturalnej jest wektor lub punkt w przestrzeni n -wymiarowej.

Standardowy podział cech służących do opisu własności struktur obiektów na mierzalne /liczbowe/ i niemierzalne /opisowe/ stanowi duże uproszczenie i dlatego istotną sprawą jest stwierdzenie ich różności pod względem stosowanych skal pomiaru.

S. S. Stevens [1, s.247] wyróżnia następujące cztery rodzaje skal pomiaru, a mianowicie : nominalną /klasyfikującą/, porządkową /rangową/, interwałową /przedziałową/ oraz ilorazową /stosunkową/. W tym miejscu zostaną omówione dwie ostatnie /najsilniejsze/ skale pomiaru, ponieważ skala nominalna oraz porządkowa w badaniach strukturalnych nie odgrywa poważniejszej roli.

Własnością cech liczbowych mierzonych na skali interwałowej lub ilorazowej jest możliwość jednoznacznego uporządkowania ich wariantów na osi liczbowej /poprzez wprowadzenie stałej jednostki miary/. Zasadniczym ograniczeniem skali interwałowej jest umowność punktu zerowego co sprawia, że niedopuszczalne jest przy tej skali mnożenie i dzielenie. W konsekwencji nie można operować stosunkami między stopniami skali. Błędne jest więc stwierdzenie, że saldo strat i zysków w przedsiębiorstwie Z_1 jest dwukrotnie większe niż w przedsiębiorstwie Z_2 . Zmiana jednostek na skali interwałowej nie zmienia proporcjonalności różnic między pomiarami. W skali tej dopuszcza się więc następujące relacje: równość - nierówność, większy - mniejszy, równość różnic bezwzględnych.

Skala ilorazowa charakteryzuje się posiadaniem absolutnego punktu zerowego /tzn. zero jednostronnie ogranicza jej zakres/. Zmiana jednostek na tej skali nie zmienia ilorazu wyników obserwacji dotyczących dowolnych, ustalonych par obiektów. Na skali tego typu dopuszcza się dodatkowo operowanie stosunkami między stopniami skali. Tylko dla cech mierzonych na skali ilorazowej można powiedzieć, że np. produkcja globalna w przedsiębiorstwie Z_1 wynosząca 4 mln zł jest dwa razy wyższa niż w przedsiębiorstwie Z_2 , gdzie

wynosi 2 mln zł.

W badaniach strukturalnych stosuje się przede wszystkim cechy mierzone w skali ilorazowej, najbardziej precyzyjnej i najmocniejszej.

Istotne znaczenie teoretyczne ma podział cech na skokowe /zbiór wartości cechy jest zbiorem przeliczalnym/ oraz ciągłe /zbiór możliwych wartości cechy jest zbiorem nieprzeliczalnym/. Przy pomiarach cech skokowych wielkości mogą zwiększać się lub zmniejszać od razu jedynie o pełną jednostkę, podczas gdy przy cechach ciągłych wzrost lub spadek może być dowolnie małą częścią jednostki. W praktyce statystycznej - jak słusznie stwierdza J. P. Guilford [36, s. 37] - "traktujemy pomiary cech ciągłych i skokowych na ogół w ten sam sposób". Wynika to z faktu, że pomiaru wartości cechy ciągłej dokonuje się zawsze z pewnym przybliżeniem.

Takie postawienie problemu pomiaru cech w badaniach strukturalnych nie powoduje żadnych ujemnych konsekwencji.

W zakresie wyboru cech w badaniach strukturalnych można wyróżnić dwa zagadnienia:

- 1^o ustalenie priorytetu cechy w klasyfikacji jednocechowej,
- 2^o wybór pewnej grupy cech w klasyfikacji wielocechowej /zwanym cechami diagnostycznymi/.

W klasyfikacji jednocechowej zakłada się, że wybrana cecha mająca charakter reprezentacyjny w odniesieniu do zbioru cech, pozwoli przeprowadzić wartościową klasyfikację z punktu widzenia celu badania. Jednostronności wyboru w klasyfikacji jednocechowej zapobiega klasyfikacja wielocechowa, która opiera się na zespole cech diagnostycznych.

Dla rozwiązania powyższych zagadnień w zakresie doboru cech służą trzy następujące podejścia:

- 1^o dobór merytoryczny w ścisłym tego słowa znaczeniu,
- 2^o dobór merytoryczno-formalny,

3^o dobór merytoryczny z pewnymi aspektami doboru formalnego.

Punktem wyjścia wszystkich trzech podejść jest skonstruowanie w oparciu o merytoryczną znajomość zagadnienia potencjalnej /wstępnej/ listy cech. Zadanie to jest szczególnie odpowiedzialne. Badaczowi nie wolno z jednej strony opuścić żadnej cechy mającej istotne znaczenie przy wyjaśnianiu przedmiotu badania, a z drugiej strony z listy należy usunąć cechy, które słabo, pośrednio lub pozornie wyjaśniają przedmiot badania. Mniejszym złem - jak z tego widać - jest wprowadzenie cech nieistotnych w początkowej fazie, niż opuszczenie cech istotnych. Cechy nieistotne mogą być w dalszej fazie usunięte, natomiast pominięcie tych drugich wypacza w dużym stopniu wyniki badań [6].

Przy konstrukcji potencjalnej listy cech trzeba mieć na względzie następujące wymogi:

- a/ postulat ekonomiczności badań; chodzi tutaj o koszt uzyskania informacji o cechach;
- b/ dostępność danych statystycznych;
- c/ wiarygodność danych statystycznych;
- d/ mierzalność cech; lista cech winna być zbiorem cech kwantytatywnych /liczbowych/; jest to warunek konieczny w badaniach strukturalnych przy użyciu metod statystyczno-matematycznych.

Tak ustalona wstępna lista cech jest - jak pisze autorka pracy [15, s.110] - "wypadkową znajomości przedmiotu badania oraz tradycji statystycznej /zbiera się dane statystyczne dotyczące dobrze już znanych zjawisk/".

Merytoryczny dobór cech jest działaniem w głównej mierze subiektywnym. Redukcji potencjalnej listy cech dokonuje się na podstawie własnej znajomości przedmiotu badania, wykorzystując współpracę przedstawicieli odpowiednich dyscyplin naukowych /ekspertów/ oraz opierając się na szeroko pojętej teorii ekonomii.

Dobry zestaw cech wybrany według kryteriów merytorycznych powinien zawierać takie, które [16; 24; 81] :

- a/ ujmują konieczne i istotne właściwości badanej zbiorowości; postuluje się aby wybrane cechy miały ustaloną interpretację i ich wskazania mogłyby podlegać merytorycznej weryfikacji;
- b/ są proste i logicznie ze sobą powiązane;
- c/ są jasno sprecyzowane, czyli ściśle zdefiniowane; cecha jest ściśle zdefiniowana wtedy, gdy nie ma różnego zakresu informacyjnego dla różnych obiektów;
- d/ wybrane cechy powinny tworzyć podzbiór nieliczny, a jednocześnie zapewniać realizację punktu a/;
- e/ zachowują swą wysoką wartość poznawczą w przyszłości, nawet wtedy, gdy badanie bezpośrednio nie służy celom prognostycznym; postulat ten opiera się na przeświadczeniu, że tylko ta analiza przeszłości jest wartościowa, która prowadzi do sformułowania wniosków ważnych dla przyszłości;
- f/ są uniwersalne, tzn. mają ustaloną wysoką wartość poznawczą w stosunku do wszystkich rozpatrywanych obiektów /nie mogą to być cechy specyficzne, a więc związane tylko z niektórymi obiektami/.

Do powyższych własności cech diagnostycznych dodaje się te, których realizacja wymaga analizy formalnej /dla której konieczne jest zgromadzenie danych statystycznych/:

- g/ mają dużą zawartość informacyjną, czyli posiadają zdolność dyskryminacji badanych jednostek;
- h/ nie powielają tych samych informacji; sytuacja taka ma miejsce wtedy, gdy cechy diagnostyczne są wzajemnie niezależne oraz gdy są one zależne od cech nie wchodzących do zespołu diagnostycznego; w niniejszym badaniu, ograniczającym się do wykrywania liniowych zależności między cechami warunek ten oznacza, że cechy są nieskorelowane lub co najwyżej słabo skorelowane między sobą oraz są silnie skorelowane z cechami nie wchodzącymi do zespołu diagnosti-

cznego [77; 79] .

Powielanie informacji zawartych w dwóch cechach można mierzyć przy pomocy:

1^o współczynnika korelacji liniowej K. Pearsona r_{11} ,

2^o miar quasi-podobieństwa cech /odpowiednio quasi-odległości i quasi-bliskości/³ wykorzystujących w swej konstrukcji współczynnik korelacji liniowej [4; 9; 34; 39] :

$$1 - |r_{11}| , \quad |r_{11}| ,$$

$$1 - r_{11}^2 , \quad r_{11}^2 , \quad /2.10/$$

$$\left[1 - |r_{11}|\right]^{0,5} , \quad \left[1 - r_{11}^2\right]^{0,5} .$$

Podjęcie drugie merytoryczno-formalne polega na tym, że ze wstępnej listy cech /wybranych na podstawie analizy merytorycznej/ usuwa się najpierw cechy, które charakteryzują się małą zawartością informacyjną /mierzona jest ona tradycyjnie zmiennością/. Usuwa się więc z listy cechy, które wykazują małą zmienność. Podstawą eliminacji cech może być np. współczynnik zmienności.⁴ W zależności od tego czy interesuje nas skala czy kształt w strukturze musimy najpierw unormować wartości cech według formuł przedstawionych w podrozdziale 2.2. Współczynnik zmienności należy ustalić dla każdego elementu struktury z osobna, a następnie z tak otrzymanych wartości policzyć średnią arytmetyczną, która będzie podstawą eliminacji danej cechy strukturalnej. Następnie do tak zredukowanej liczby cech stosuje się formalny algorytm wyboru cech diagnostycznych, który zapewnia realizację własności h/ dobrego zestawu cech. Dobry przegląd formalnych metod doboru cech znajduje się w pracach [34; 93].

³ Miary quasi-podobieństwa cech nie spełniają warunku 2^o określonego w rozdziale trzecim dla miar podobieństwa.

⁴ Inne miary zmienności omówione są m. in. w pracy [93] .

W związku z tym, że silne skorelowanie cech nie zawsze oznacza powielanie przez nie informacji /z punktu widzenia badanego problemu mogą nieść one inne informacje/ stosuje się podejście trzecie oparte na doborze merytorycznym z pewnymi aspektami doboru formalnego . Podejście to cechuje się jednak dużą dozą subiektywizmu. Można je streścić w postaci następujących punktów:

- a/ ze wstępnej listy cech strukturalnych zostają usunięte te, które wykazują małą zmienność;⁵
- b/ na podstawie analizy merytorycznej pozostałych cech wyodrębnione zostają grupy cech, co do których możemy przypuszczać iż mogą powielać informacje;
- c/ dla każdej grupy cech wyodrębnionej w punkcie b/ stosujemy formalny algorytm doboru cech diagnostycznych.

Podejście powyższe jest pracochłonne i wymaga od badacza bardzo dobrej znajomości problematyki badawczej.

Analizując zaprezentowane podejścia w zakresie doboru cech autor uważa, że właśnie ostatnie z nich wydaje się być najbardziej efektywne i należałoby stosować je wtedy, gdy zezwala na to posiadany materiał empiryczny.

2.2. Prezentacja wielkości /skali/ oraz kształtu /formy/ struktur.

Otrzymana macierz danych o zapisie /2.5/ stanowi etap wstępny do prezentacji wielkości /skali/ oraz kształtu /formy/ struktur.⁶ Specyfika tych zagadnień wymaga odmiennego unormowania cech strukturalnych.⁷

⁵ Uwagi odnośnie ustalania zmienności cech jak przy doborze merytoryczno-formalnym.

⁶ Jest to skrót myślowy, bowiem mowa będzie o prezentacji wielkości /skali/ oraz kształtu /formy/ obrazów struktur obiektów.

⁷ Omawiane reguły normowania przystosowane są dla cech mierzonych na skali ilorazowej.

Zanim zostaną przedstawione sposoby normowania cech strukturalnych należy określić warunki, jakie powinny one spełniać, aby umożliwiały prezentację kształtu oraz skali struktur.

Poprawna reguła normowania cech strukturalnych umożliwiająca prezentację kształtu powinna spełniać taki warunek, że jeśli kąty między odpowiadającymi sobie wektorami kolumnowymi w macierzach $X_{..r}$ i $X_{..s}$ są równe 0° , to dwie struktury obiektów, tj. $S/A_r/$ i $S/A_s/$ prezentują identyczny kształt.⁸ Natomiast będą one prezentować zbliżony kształt wtedy, gdy kąty te nieznacznie będą różnić się od 0° .

Z kolei poprawna reguła normowania cech strukturalnych umożliwiająca prezentację skali powinna spełniać taki warunek, że jeśli poszczególne składowe odpowiadających sobie kolumn w macierzach $X_{..r}$ i $X_{..s}$ przyjmują te same wartości, to struktury obiektów $S/A_r/$ i $S/A_s/$ prezentują identyczną skalę. Inaczej mówiąc dwie struktury obiektów, tj. $S/A_r/$ i $S/A_s/$ prezentują identyczną skalę wtedy i tylko wtedy, gdy kąty między odpowiadającymi sobie wektorami kolumnowymi w macierzach $X_{..r}$ i $X_{..s}$ są równe 0° oraz długości odpowiadających sobie wektorów są jednakowe. Będą one prezentować zbliżoną skalę wtedy, gdy kąty te nieznacznie będą różniły się od 0° i długości odpowiadających sobie wektorów będą podobne.

Tradycyjnymi sposobami normowania⁹ cech strukturalnych /dla ustalonej l-tej cechy/ są następujące formuły:

$$p_{i.r} = \frac{x_{i.r}}{\sum_{j=1}^k x_{j.r}}, \quad \begin{array}{l} i = 1, 2, \dots, k; \\ r = 1, 2, \dots, n. \end{array} \quad /2.11/$$

⁸ Kąt między dwoma wektorami jest równy 0° wtedy, gdy ich iloczyn skalarny jest równy iloczynowi ich długości.

⁹ W sensie statystycznym.

$$p_{i.r} = \frac{x_{i.r}}{\sum_{s=1}^n \sum_{j=1}^k x_{j.s}}, \quad \begin{array}{l} i = 1, 2, \dots, k; \\ r = 1, 2, \dots, n. \end{array} \quad /2.12/$$

Reguła normowania o postaci /2.11/ służy do prezentacji kształtu /formy/ struktur, natomiast formuła /2.12/ umożliwia prezentację wielkości /skali/ struktur o zapisie /2.5/. Reguły normowania /2.11/ i /2.12/ powodują, że każdy element macierzy danych po unormowaniu jest zawarty w przedziale domkniętym $\langle 0; 1 \rangle$ oraz spełnione są dla nich odpowiednio następujące relacje:

$$\sum_{i=1}^k p_{i.r} = 1, \quad \sum_{r=1}^n \sum_{i=1}^k p_{i.r} = 1. \quad /2.13 - 2.14/$$

W niniejszej pracy proponuje się [109] zastosowanie innych formuł normowania. Można je ująć w postaci następujących wzorów /dla ustalonej l-tej cechy/:

$$p_{i.r} = \frac{x_{i.r}}{\max_j x_{j.r}}, \quad \begin{array}{l} i = 1, 2, \dots, k; \\ r = 1, 2, \dots, n. \end{array} \quad /2.15/$$

$$p_{i.r} = \frac{x_{i.r}}{\max_s \max_j x_{j.s}}, \quad \begin{array}{l} i = 1, 2, \dots, k; \\ r = 1, 2, \dots, n. \end{array} \quad /2.16/$$

Formuła /2.15/ służy do prezentacji kształtu /formy/ struktur, natomiast reguła normowania /2.16/ umożliwia z kolei prezentację wielkości /skali/ struktur o zapisie /2.5/. Wzory o postaci /2.15/ i /2.16/ powodują unormowanie elementów macierzy danych w przedziale domkniętym $\langle 0; 1 \rangle$.

Przez zaprezentowanie autorskich propozycji normowania cech

strukturalnych, umożliwiającą prezentację kształtu oraz skali struktur, wykazano że istnieją również inne niż tradycyjne sposoby normowania. Proponowane reguły pozwalają osiągnąć ten sam cel badawczy co formuły /2.11/ i /2.12/. Z punktu widzenia badań strukturalnych formuły /2.15/ i /2.16/ można uznać za równoważne formułom /2.11/ i /2.12/.

W badaniach wielocechowych reguły normowania umożliwiają dodatkowo sprowadzenie cech różnoimiennych /wyrażonych w różnych jednostkach miary/ do porównywalności.

Zastosowanie jednej z formuł normowania zaprezentowanych powyżej powoduje przekształcenie blokowej macierzy danych o zapisie /2.5/ w macierz blokową o unormowanych wartościach cech strukturalnych. Unormowany obraz struktury r-tego obiektu przedstawia macierz:

$$\begin{bmatrix} p_{11r} & p_{12r} & \dots & p_{1mr} \\ p_{21r} & p_{22r} & \dots & p_{2mr} \\ \dots & \dots & \dots & \dots \\ p_{k1r} & p_{k2r} & \dots & p_{kmr} \end{bmatrix} = \mathbb{P}_{..r} \quad /2.17/$$

gdzie: p_{ilr} - unormowana wartość i-tego elementu struktury zmierzona dla l-tej cechy w r-tym obiekcie;

$\mathbb{P}_{..r}$ - unormowany obraz struktury obiektu A_r ;

pozostałe oznaczenia bez zmian.

W przypadku, gdy wybrane zjawisko /problem badawczy/ z zakresu struktur obiektów będzie rozpatrywane ze względu na jego poziom, to należy stosować reguły normowania umożliwiające prezentację skali struktur. Jeśli wybrane zjawisko będzie rozpatrywane ze względu na jego profil /typ/, to należy stosować reguły normowania umożliwiające prezentację kształtu struktur.

3. MIARY PODOBIENSTWA STRUKTUR

3.1. Uwagi wstępne.

Aby można było tworzyć klasy, konieczne jest wypracowanie sposobu, który umożliwi zaliczanie obiektów podobnych do jednej klasy. Zatem podstawowym problemem, na tym etapie metod klasyfikacji, jest formalizacja pojęcia "podobieństwo". Stopień podobieństwa obiektów kwantyfikuje się za pomocą miar podobieństwa, wśród których wyróżnia się miary odległości oraz bliskości [19; 85].

Funkcja

$$d : A \times A \rightarrow R \text{ /zbiór liczb rzeczywistych/}$$

nazywana będzie miarą odległości wtedy i tylko wtedy, gdy spełnione będą następujące warunki:

- 1° $\begin{array}{c} \wedge \\ A_r, A_s \in A \end{array}$ $d/A_r, A_s/ \geq 0$ - warunek nieujemności;
- 2° $\begin{array}{c} \wedge \\ A_r, A_s \in A \end{array}$ $d/A_r, A_s/ = 0 \iff r = s$ - warunek zwrotności;
- 3° $\begin{array}{c} \wedge \\ A_r, A_s \in A \end{array}$ $d/A_r, A_s/ = d/A_s, A_r/$ - warunek symetryczności.

Jeżeli będzie spełniony ponadto warunek trójkąta:

4° $\begin{array}{c} \wedge \\ A_r, A_s, A_t \in A \end{array}$ $d/A_r, A_s/ \leq d/A_r, A_t/ + d/A_t, A_s/$

to miara odległości zwana jest metryką.

O ile warunki 1° - 3° muszą być spełnione /mają charakter obligatoryjny/, o tyle warunek 4° jest niekonieczny, tzn. ma charakter fakultatywny.

Autor do powyższych warunków dodaje jeszcze jeden, który jego zdaniem w badaniach strukturalnych powinien być obligatoryjny.

Jeśli przy warunku 2^o określona została dolna granica miary odległości /porównywane obiekty są wtedy identyczne¹/, to również należy określić górną granicę, która oznacza, że porównywane obiekty są zupełnie niepodobne. Warunek ten przyjmuje następujący zapis:

$$5^o \quad \begin{array}{c} \wedge \\ A_r, A_s \in A \end{array} \quad d/A_r, A_s/ \leq 1,$$

$$\begin{array}{c} \wedge \\ A_r, A_s \in A \end{array} \quad d/A_r, A_s/ = 1. \text{ Sytuacja taka ma miejsce wtedy,}$$

gdy nie ma ani jednego elementu struktury, który jednocześnie reprezentowany jest w strukturze obiektu A_r i A_s , tzn. dla każdego j zachodzi jedna z następujących sytuacji:

$$a/ K_j/A_r/ = \emptyset \wedge K_j/A_s/ = \emptyset ;$$

$$b/ K_j/A_r/ = \emptyset \wedge K_j/A_s/ \neq \emptyset ;$$

$$c/ K_j/A_r/ \neq \emptyset \wedge K_j/A_s/ = \emptyset .$$

Na analogicznych zasadach określona zostanie w tym miejscu funkcja bliskości.

Funkcja

$$g : A \times A \text{ ----} \rightarrow R$$

nazywana będzie miarą bliskości wtedy i tylko wtedy, gdy spełnione będą następujące warunki:

$$1^o \quad \begin{array}{c} \wedge \\ A_r, A_s \in A \end{array} \quad 0 \leq g/A_r, A_s/ \leq 1 ;$$

$$2^o \quad \begin{array}{c} \wedge \\ A_r, A_s \in A \end{array} \quad g/A_r, A_s/ = 1 \iff r = s ;$$

$$3^o \quad \begin{array}{c} \wedge \\ A_r, A_s \in A \end{array} \quad g/A_r, A_s/ = g/A_s, A_r/ .$$

¹Tylko w sytuacji, gdy porównywane obiekty są tym samym obiektem.

Jeżeli ponadto spełniony będzie warunek trójkąta:

$$4^{\circ} \quad \begin{array}{c} \wedge \\ A_r, A_s, A_t \in A \end{array} \quad g/A_{r,A_t}/ \geq \frac{g/A_{r,A_s}/ \cdot g/A_{s,A_t}/}{g/A_{r,A_s}/ + g/A_{s,A_t}/}$$

to miara bliskości zwana jest metryką.

O ile warunki 1° - 3° mają charakter obligatoryjny, o tyle warunek 4° jest niekonieczny, tzn. ma charakter fakultatywny.

Podobnie jak przy mierze odległości również tutaj zostanie wprowadzony dodatkowy warunek. Jeśli przy warunku 2° określona została górna granica miary bliskości /porównywane obiekty są wtedy identyczne/, to również należy określić granicę dolną, która oznacza, że porównywane obiekty są zupełnie niepodobne. Przyjmuje on następującą postać:

$$5^{\circ} \quad \begin{array}{c} \wedge \\ A_r, A_s \in A \end{array} \quad g/A_{r,A_s}/ = 0. \text{ Opis sytuacji w której reali-}$$

zuje się ten warunek jest identyczny z opisem zamieszczonym w warunku 5° przy omawianiu miary odległości.

Miary odległości można przy pomocy odpowiednich wzorów przekształcić na miary bliskości i odwrotnie /por. prace [18; 98] /. Takie "naturalne" przekształcenie wyraża wzór:

$$d/A_{r,A_s}/ = 1 - g/A_{r,A_s}/ . \quad /3.1/$$

Miary podobieństwa obiektów A_r i A_s obliczane są na podstawie unormowanych obrazów struktur tych obiektów, tj. na podstawie macierzy $\mathbb{P}_{..r}$ i $\mathbb{P}_{..s}$. W związku z tym miary te oznaczane będą następująco:

$$d/A_{r,A_s}/ = d[S/A_r/; S/A_s/] = d/\mathbb{P}_{..r}; \mathbb{P}_{..s}/ = d_{rs} - \text{miary odległości;}$$

$$g/A_r, A_s/ = g[S/A_r/; S/A_s/] = g/P_{..r}; P_{..s}/ = g_{rs} - \text{miary bliskości.}$$

Szczegółowego wyjaśnienia wymagają więc niektóre z postawionych warunków dla miar odległości i bliskości. Warunek 2^o będzie interpretowany następująco: dwa obiekty A_r i A_s są "identyczne" wtedy, gdy odpowiadające sobie składowe w unormowanych macierzach obserwacji $P_{..r}$ i $P_{..s}$ będą sobie równe. Widać z tego, że ustalanie podobieństwa obiektów A_r i A_s na podstawie unormowanych obrazów struktur tych obiektów powoduje sytuację, w której identyczność między dwoma obiektami może zachodzić nie tylko wtedy, gdy są one tym samym obiektem. Z kolei z warunku 5^o wynika, że dwa obiekty A_r i A_s są zupełnie niepodobne wtedy, gdy iloczyn macierzy $P_{..r}^T P_{..s}$ jest macierzą, w której na głównej przekątnej są same zera. Inaczej mówiąc odpowiadające sobie wektory kolumnowe w macierzach obserwacji $P_{..r}$ i $P_{..s}$ są ortogonalne.²

W literaturze można spotkać bardzo dużo propozycji miar podobieństwa. Dobry przegląd niektórych typów miar można znaleźć w pracach R. M. Cormacka [17], P. H. A. Sneatha i R. R. Sokala [89], M. R. Anderberga [4], C. R. Rao [84]. Jednakże z punktu widzenia warunków nałożonych w pracy na miary podobieństwa liczba ich staje się znacznie ograniczona. Wśród tych, które spełniają te warunki są miary posiadające pewne wady z punktu widzenia badań strukturalnych oraz miary, których zastosowanie ogranicza się tylko do badania podobieństwa kształtu. W związku z tym autor niniejszej pracy uznaje za celowe podanie nowych propozycji w tym zakresie.

Wszystkie zaprezentowane w tym rozdziale miary podobieństwa struktur³ dotyczyć będą struktur prostych /zbiór cech jest jedno-

²Wektory są ortogonalne, jeśli ich iloczyn skalarny równa się zero.

³Jest to skrót myślowy. Mowa będzie o miarach podobieństwa struktur obiektów obliczanych na podstawie ich unormowanych obrazów.

elementowy/. Wprowadzając do każdego wzoru dodatkowe sumowanie po cechach /oraz indeks l w miejsce kropki/, czyli dołączając następujący człon

$$\frac{1}{m} \sum_{l=1}^m \quad , \quad /3.2/$$

otrzymuje się miary podobieństwa struktur złożonych /zbiór cech jest co najmniej dwuelementowy/.

3.2. Przegląd miar podobieństwa struktur.

Wszystkie miary odległości oraz bliskości, które będą zaprezentowane w tym rozdziale mają analogiczną interpretację /choć ze względu na odmienne konstrukcje przyjmują na ogół różne wartości liczbowe/. Interpretacja miar odległości jest następująca: w miarę jak ich wartości dążą do zera oznacza to coraz mniejsze zróżnicowanie struktur badanych obiektów, podczas gdy wzrost wartości do jedności oznacza coraz to większe różnice w strukturach porównywanych obiektów. Miary bliskości posiadają interpretację odwrotną w stosunku do miar odległości.

Wśród miar podobieństwa wyróżnia się dwie grupy. Pierwsza z nich obejmuje miary, które mają ograniczony zakres, tzn. mogą być zastosowane tylko przy badaniu podobieństwa kształtu struktur i to przy wykorzystaniu wyłącznie formuły /2.11/ normowania cech strukturalnych.

Funkcję cosinus do badań strukturalnych zastosował W. Borysiuk w pracy [10]. Wykorzystał on tę własność, że obrazy struktur obiektów przedstawiane są w praktyce w postaci wektorów w przestrzeni k -wymiarowej. Podobieństwo w sensie bliskości można ustalić obliczając cosinus kąta zawartego między nimi:

$$g_{rs} /1/ = \frac{\sum_{i=1}^k p_{i.r} p_{i.s}}{\left(\sum_{i=1}^k p_{i.r}^2\right)^{0,5} \left(\sum_{i=1}^k p_{i.s}^2\right)^{0,5}} \quad , \quad /3.3/$$

Innego typu miarą bliskości jest współczynnik podobieństwa [13] o postaci:

$$g_{rs}^{/2/} = \sum_{i=1}^k \min /p_{i.r}; p_{i.s}/ \quad /3.4/$$

Historycznie rzecz biorąc najstarszą miarą odległości struktur obiektów jest propozycja E. Hellingera [84] z 1909 r., w której wykorzystuje się funkcję odwrotną do funkcji cosinus:

$$d_{rs}^{/1/} = \frac{2}{\pi} \arccos \sum_{i=1}^k /p_{i.r} p_{i.s}/^{0,5} \quad /3.5/$$

Bhattacharya zaproponował w 1946 r. [84] miarę odległości, która w niniejszej pracy zostanie podana w wersji unormowanej w przedziale $\langle 0; 1 \rangle$:

$$d_{rs}^{/2/} = \left(1 - \sum_{i=1}^k \sqrt{p_{i.r} p_{i.s}} \right)^{0,5} \quad /3.6/$$

Drugą grupę miar podobieństwa stanowią miary odległości, które mogą być stosowane zarówno przy badaniu podobieństwa wielkości /skali/ jak i kształtu /formy/ struktur, przy wykorzystaniu formuł normowania cech strukturalnych przedstawionych w rozdziale drugim.

Pierwsze dwie miary, a mianowicie miara zwana odległością Canberra [89]

$$d_{rs}^{/3/} = \frac{1}{k} \sum_{i=1}^k \frac{|p_{i.r} - p_{i.s}|}{/p_{i.r} + p_{i.s}/} \quad /3.7/$$

oraz współczynnik dywergencji Clarka [89]

$$d_{rs}^{/4/} = \left[\frac{1}{k} \sum_{i=1}^k \left(\frac{p_{i.r} - p_{i.s}}{p_{i.r} + p_{i.s}} \right)^2 \right]^{0,5} \quad /3.8/$$

spełniają warunki obligatoryjne określone w punkcie pierwszym tego rozdziału, jednakże posiadają pewne wady. Miary tego typu nie mogą być zastosowane w badaniach strukturalnych wtedy, gdy w porównywanych strukturach obiektów nie są jednocześnie reprezentowane dla 1-tej cechy niektóre elementy struktury. Drugą wadą wynika z faktu, że typ danych niekiedy wypacza wyniki obliczeń otrzymane za pomocą tych miar. Sytuację taką najlepiej jest przeanalizować na przykładzie. Niech dane będą unormowane obrazy czterech struktur obiektów, tj.:

$$P_{..1} = /0 \ 0 \ 0 \ 0 \ 1/^T ,$$

$$P_{..2} = /0,01 \ 0,01 \ 0,01 \ 0,01 \ 0,96/^T ,$$

$$P_{..3} = /0,40 \ 0,20 \ 0,20 \ 0,10 \ 0,10/^T ,$$

$$P_{..4} = /0,38 \ 0,18 \ 0,18 \ 0,08 \ 0,10/^T .$$

Odległości między $P_{..1}$ i $P_{..2}$ oraz $P_{..3}$ i $P_{..4}$ policzone zarówno w oparciu o miarę $d_{rs}^{/3/}$ jak i $d_{rs}^{/4/}$ /por. tab. 3.1/ powinny mieć zbliżoną lub identyczną wartość /tak reagują prawie wszystkie miary omawiane w tym rozdziale/. Z tabeli widać, że porównywane odległości między $P_{..1}$ i $P_{..2}$ oraz $P_{..3}$ i $P_{..4}$, obliczone zarówno według miary /3.7/ jak i /3.8/, są diametralnie różne. Wady te w poważnym stopniu ograniczają zastosowanie tych miar w badaniach strukturalnych.

Tab. 3.1

Odległość między	Miara o postaci	
	/3.7/	/3.8/
$P_{..1}$ i $P_{..2}$	0,80	0,89
$P_{..3}$ i $P_{..4}$	0,04	0,06

Zródło: obliczenia własne.

Z tej grupy omówione zostaną jeszcze dwie miary odległości. Pierwszą z nich zaproponowali Bray i Curtis w 1957 r. [17] :

$$d_{rs}^{/5/} = \frac{\sum_{i=1}^k |p_{i.r} - p_{i.s}|}{\sum_{i=1}^k (p_{i.r} + p_{i.s})} \quad /3.9/$$

Miara ta przy badaniu podobieństwa kształtu w oparciu o formułę /2.11/ normowania cech strukturalnych redukuje się do postaci [14; 51] :

$$d_{rs}^{/6/} = \frac{1}{2} \sum_{i=1}^k |p_{i.r} - p_{i.s}| \quad /3.10/$$

Drugą propozycję podali E. Marczewski i H. Steinhaus w pracy [60] :

$$d_{rs}^{/7/} = \frac{\sum_{i=1}^k |p_{i.r} - p_{i.s}|}{\sum_{i=1}^k \max(p_{i.r}; p_{i.s})} \quad /3.11/$$

Powyższe miary nie posiadają wad miar /3.7/ i /3.8/, a co za tym idzie są godne polecenia w badaniach strukturalnych.

Między miarami /3.9/ i /3.11/ występuje następująca zależność:

$$d_{rs}^{/5/} \leq d_{rs}^{/7/} \quad /3.12/$$

Nierówność ta będzie prawdziwa, jeżeli:

$$\sum_{i=1}^k \max(p_{i.r}; p_{i.s}) \leq \sum_{i=1}^k (p_{i.r} + p_{i.s})$$

Należy wykazać, że

$$\bigwedge_i \max(p_{i.r}; p_{i.s}) \leq p_{i.r} + p_{i.s}$$

$$\max /p_{i.r}; p_{i.s}/ \leq \max /p_{i.r}; p_{i.s}/ + \min /p_{i.r}; p_{i.s}/,$$

$$\min /p_{i.r}; p_{i.s}/ \geq 0.$$

Zważywszy na to, że $p_{i.r}$ oraz $p_{i.s}$ zawierają się w przedziale $\langle 0; 1 \rangle$, przeto nierówność powyższa jest prawdziwa.

3.3. Propozycje miar podobieństwa struktur.

Na początku niniejszego rozdziału stwierdzono zasadność podania nowych propozycji miar podobieństwa struktur. Wszystkie miary przedstawione w tym podrozdziale mogą być stosowane zarówno przy badaniu podobieństwa skali jak i kształtu struktur, przy wykorzystaniu formuł normowania cech strukturalnych przedstawionych w rozdziale drugim. Najpierw zaprezentowane zostaną trzy typy rodzin miar odległości, które spełniają warunki obligatoryjne określone w punkcie 3.1 [108; 111; 112] :

$$d_{rs}^{/8/} = \left[\frac{\sum_{i=1}^k |p_{i.r} - p_{i.s}|^w}{\sum_{i=1}^k /p_{i.r} + p_{i.s}/^w} \right]^{\frac{1}{w}}, \quad /3.13/$$

$$d_{rs}^{/9/} = \left[\frac{\sum_{i=1}^k |p_{i.r}^w - p_{i.s}^w|}{\sum_{i=1}^k /p_{i.r} + p_{i.s}/^w} \right]^{\frac{1}{w}}, \quad /3.14/$$

$$d_{rs}^{/10/} = \frac{\sum_{i=1}^k |p_{i.r}^w - p_{i.s}^w|}{\sum_{i=1}^k /p_{i.r} + p_{i.s}/^w}, \quad /3.15/$$

gdzie: $w \in \langle 1, 2, \dots, \infty \rangle$.

W praktyce za w przyjmuje się jedną z dwóch liczb, tzn. 1 lub 2. Po podstawieniu za w liczby 1 otrzymuje się bardzo ciekawą sytuację. Otóż miary o numerach /3.13/ - /3.15/ przyjmują identyczną postać, którą zaprezentowano w podrozdziale 3.2. Jest to miara Bray'a i Curtisa o zapisie /3.9/.

Zasadnicze różnice występują po podstawieniu za w liczby 2:

$$d_{rs}^{/11/} = \left[\frac{\sum_{i=1}^k /p_{i.r} - p_{i.s}/^2}{\sum_{i=1}^k /p_{i.r} + p_{i.s}/^2} \right]^{0,5} \quad /3.16/$$

$$d_{rs}^{/12/} = \left[\frac{\sum_{i=1}^k |p_{i.r}^2 - p_{i.s}^2|}{\sum_{i=1}^k /p_{i.r} + p_{i.s}/^2} \right]^{0,5} \quad /3.17/$$

$$d_{rs}^{/13/} = \frac{\sum_{i=1}^k |p_{i.r}^2 - p_{i.s}^2|}{\sum_{i=1}^k /p_{i.r} + p_{i.s}/^2} \quad /3.18/$$

Między miarami o numerach /3.16/ - /3.18/ zachodzą następujące relacje:

$$d_{rs}^{/13/} \leq d_{rs}^{/11/} \leq d_{rs}^{/12/} \quad /3.19/$$

Pierwsza nierówność w relacji /3.19/ jest następująca:

$$\frac{\sum_{i=1}^k |p_{i.r}^2 - p_{i.s}^2|}{\sum_{i=1}^k /p_{i.r} + p_{i.s}/^2} \leq \left[\frac{\sum_{i=1}^k /p_{i.r} - p_{i.s}/^2}{\sum_{i=1}^k /p_{i.r} + p_{i.s}/^2} \right]^{0,5} \cdot /3.20/$$

Po podniesieniu do kwadratu obu stron nierówności i przemnożeniu ich przez

$$\left(\sum_{i=1}^k /p_{i.r} + p_{i.s}/^2 \right)^2$$

należy przenieść prawą stronę na lewą otrzymując:

$$\left(\sum_{i \in K} |p_{i.r}^2 - p_{i.s}^2| \right)^2 - \sum_{i \in K} /p_{i.r} - p_{i.s}/^2 \sum_{i \in K} /p_{i.r} + p_{i.s}/^2 \leq 0.$$

Celem eliminacji wartości bezwzględnej zbiór K podzielono na dwa podzbiory K_1 i K_2 takie, że

i-te elementy dla których $p_{i.r} \geq p_{i.s}$ tworzą zbiór K_1 , a

i-te elementy dla których $p_{i.r} < p_{i.s}$ tworzą zbiór K_2 :

$$\left(\sum_{i \in K_1} /p_{i.r} - p_{i.s}/^2 + \sum_{i \in K_2} /p_{i.s} - p_{i.r}/^2 \right)^2 - \left(\sum_{i \in K_1} /p_{i.r} - p_{i.s}/^2 + \sum_{i \in K_2} /p_{i.r} - p_{i.s}/^2 \right) \left(\sum_{i \in K_1} /p_{i.r} + p_{i.s}/^2 + \sum_{i \in K_2} /p_{i.r} + p_{i.s}/^2 \right) \leq 0.$$

Po wykonaniu wszystkich czynności po lewej stronie nierówności oraz po dokonaniu redukcji wyrazów podobnych otrzymano:

$$\left(\sum_{i \in K_1} p_{i.r} p_{i.s} + \sum_{i \in K_2} p_{i.r} p_{i.s} \right)^2 \leq \left(\sum_{i \in K_1} p_{i.r}^2 + \sum_{i \in K_2} p_{i.s}^2 \right)$$

$$\left(\sum_{i \in K_1} p_{i.s}^2 + \sum_{i \in K_2} p_{i.r}^2 \right) ,$$

$$\left(\sum_{i \in K} p_{i.r} p_{i.s} \right)^2 \leq \left(\sum_{i \in K} \max /p_{i.r}; p_{i.s}' \right)^2 \left(\sum_{i \in K} \min /p_{i.r}; p_{i.s}' \right)^2 ,$$

$$\left(\sum_{i \in K} \min /p_{i.r}; p_{i.s}' / \max /p_{i.r}; p_{i.s}' \right)^2 \leq$$

$$\left(\sum_{i \in K} \max /p_{i.r}; p_{i.s}' \right)^2 \left(\sum_{i \in K} \min /p_{i.r}; p_{i.s}' \right)^2 .$$

Powyższa postać zwana jest nierównością Schwarz-Cauchy'ego. Oczywiście na mocy prawdziwości tej nierówności [88] wykazano prawdziwość nierówności o zapisie /3.20/.

Z relacji /3.19/ pozostała do udowodnienia nierówność druga, tzn.:

$$\left[\frac{\sum_{i=1}^k /p_{i.r} - p_{i.s}'/2}{\sum_{i=1}^k /p_{i.r} + p_{i.s}'/2} \right]^{0,5} \leq \left[\frac{\sum_{i=1}^k |p_{i.r}^2 - p_{i.s}'^2|}{\sum_{i=1}^k /p_{i.r} + p_{i.s}'/2} \right]^{0,5} \quad . /3.21/$$

Nierówność ta będzie prawdziwa, jeżeli:

$$\sum_{i=1}^k /p_{i.r} - p_{i.s}'/2 \leq \sum_{i=1}^k |p_{i.r}^2 - p_{i.s}'^2| .$$

Należy wykazać, że

$$\bigwedge_i /p_{i.r} - p_{i.s}'/2 \leq |p_{i.r}^2 - p_{i.s}'^2| ,$$

$$|p_{i.r}^2 - p_{i.s}^2| = \begin{cases} /a/ & \left\{ \begin{array}{l} p_{i.r}^2 - p_{i.s}^2 \Leftrightarrow p_{i.r}^2 \gg p_{i.s}^2 \Rightarrow p_{i.r} \gg p_{i.s} \\ /b/ & \left\{ \begin{array}{l} -/p_{i.r}^2 - p_{i.s}^2/ \Leftrightarrow p_{i.r}^2 < p_{i.s}^2 \Rightarrow p_{i.r} < p_{i.s} \end{array} \right. \end{array} \right. \end{cases}$$

$$/a/ \quad \bigwedge_i |p_{i.r} - p_{i.s}|^2 \leq p_{i.r}^2 - p_{i.s}^2,$$

$$p_{i.r}^2 - 2 p_{i.r} p_{i.s} + p_{i.s}^2 \leq p_{i.r}^2 - p_{i.s}^2,$$

$$2 p_{i.s}^2 - 2 p_{i.r} p_{i.s} \leq 0 \quad /: 2 p_{i.s},$$

$$p_{i.s} - p_{i.r} \leq 0 \Rightarrow p_{i.s} \leq p_{i.r};$$

$$/b/ \quad \bigwedge_i |p_{i.r} - p_{i.s}|^2 \leq -/p_{i.r}^2 - p_{i.s}^2/,$$

$$p_{i.r}^2 - 2 p_{i.r} p_{i.s} + p_{i.s}^2 \leq -p_{i.r}^2 + p_{i.s}^2,$$

$$2 p_{i.r}^2 - 2 p_{i.r} p_{i.s} \leq 0 \quad /: 2 p_{i.r},$$

$$p_{i.r} - p_{i.s} \leq 0 \Rightarrow p_{i.r} \leq p_{i.s} \quad \text{cbdo.}$$

W tym miejscu przedstawione zostaną miary bliskości spełniające warunki obligatoryjne określone dla tego typu miar w podrozdziale 3.1.

Pierwsza propozycja autora, pokazana wcześniej w pracach [107; 110], przedstawia się następująco:

$$g_{rs}^{/3/} = \frac{1}{k - a_1} \sum_{i=1}^k \frac{\min\{p_{i.r}; p_{i.s}\}}{\max\{p_{i.r}; p_{i.s}\}}, \quad /3.22/$$

gdzie: a_1 - liczba składowych 1-tej cechy strukturalnej dla których w r-tym i s-tym obiekcie zachodzi:

$$p_{i.r} = 0 \wedge p_{i.s} = 0 .$$

Miara tego typu posiada tę samą wadę co miary /3.7/ i /3.8/ omówione w podrozdziale 3.2.

Kolejne dwie propozycje miar bliskości - wolne od wad poprzedniej - przyjmują następujące postacie:

$$g_{rs}^{/4/} = \frac{\sum_{i=1}^k \min /p_{i.r}; p_{i.s}/}{\frac{1}{2} \sum_{i=1}^k /p_{i.r} + p_{i.s}/} \quad , \quad /3.23/$$

$$g_{rs}^{/5/} = \frac{\sum_{i=1}^k \min /p_{i.r}; p_{i.s}/}{\sum_{i=1}^k \max /p_{i.r}; p_{i.s}/} \quad . \quad /3.24/$$

Można wykazać, że między miarami o zapisie /3.23/ i /3.24/ zachodzi następująca zależność:

$$g_{rs}^{/5/} \leq g_{rs}^{/4/} \quad . \quad /3.25/$$

Należy udowodnić, że

$$\sum_{i=1}^k \max /p_{i.r}; p_{i.s}/ \geq \frac{1}{2} \sum_{i=1}^k /p_{i.r} + p_{i.s}/ \quad ,$$

$$2 \sum_{i=1}^k \max /p_{i.r}; p_{i.s}/ \geq \sum_{i=1}^k \left(\min /p_{i.r}; p_{i.s}/ + \max /p_{i.r}; p_{i.s}/ \right)$$

$$2 \sum_{i=1}^k \max /p_{i.r}; p_{i.s}/ \gg \sum_{i=1}^k \min /p_{i.r}; p_{i.s}/ + \sum_{i=1}^k \max /p_{i.r}; p_{i.s}/,$$
$$\sum_{i=1}^k \max /p_{i.r}; p_{i.s}/ \gg \sum_{i=1}^k \min /p_{i.r}; p_{i.s}/ ,$$

ponieważ $\bigwedge_i \max /p_{i.r}; p_{i.s}/ \gg \min /p_{i.r}; p_{i.s}/$, więc nierówność o zapisie /3.25/ jest prawdziwa.

W związku z tym, że przyjęta w badaniu miara podobieństwa może mieć istotne znaczenie dla klasyfikacji obiektów pojawia się problem jej wyboru. Jest on o tyle skomplikowany, że nie ma możliwości formalnej oceny każdej z miar [21; 89]. Powoduje to, że w pewnym sensie wybór jednej z nich jest wyborem arbitralnym.

Można jednak podać kilka praktycznych zaleceń, które wprowadzicie nie rozwiązują problemu wyboru miary podobieństwa, ale znacznie ograniczają liczbę możliwych do zastosowania miar w badaniach strukturalnych:

1^o Miara podobieństwa powinna być nieskomplikowana formalnie, ażeby nie sprawiała trudności numerycznych. Występują one przy posługiwaniu się miarami, które wykorzystują w swojej konstrukcji funkcje trygonometryczne, potęgowanie, pierwiastkowanie. Konstrukcja miar podobieństwa nie ma praktycznie znaczenia przy posługiwaniu się elektroniczną techniką obliczeniową.

2^o Miara podobieństwa powinna być na tyle uniwersalna, żeby mogła znaleźć zastosowanie zarówno przy badaniu podobieństwa kształtu /formy/ jak i wielkości /skali/ struktur. Miary o numerach /3.3/ - /3.6/ nie spełniają tego postulatu.

3^o Typ danych nie powinien wypaczać rezultatów otrzymanych w wyniku zastosowania określonych miar podobieństwa. Miary o numerach /3.7/, /3.8/ i /3.22/ posiadają tę wadę.

Pozostałe miary podobieństwa, nie wymienione w punktach 2^o i 3^o, można uznać za ekwiwalentne jeśli chodzi o ich możliwość zastosowania w badaniach strukturalnych.

Należy mieć na uwadze to, że nie można zmieniać raz przyjętej miary podobieństwa w kolejnych fazach badania /np. macierz podobieństw musi być ustalona przy pomocy tej samej miary podobieństwa chcąc porównać zgodność wyników dwóch metod klasyfikacji/.

W oparciu o wybraną formułę miary podobieństwa należy wyliczyć symetryczną macierz podobieństw o postaci:

$$[C] = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{bmatrix}$$

/3.26/

gdzie: c_{rs} - wartość miary podobieństwa obiektów A_r i A_s obliczona na podstawie unormowanych obrazów struktur tych obiektów, tj. na podstawie macierzy $\overline{P}_{\dots r}$ i $\overline{P}_{\dots s}$;

$c_{rs} = d_{rs}$ - dla miar odległości;

$c_{rs} = g_{rs}$ - dla miar bliskości.

4. METODY KLASYFIKACJI

4.1. Wprowadzenie.

W literaturze przedmiotu oprócz pojęcia wymienionego w tytule pracy istnieje wiele innych, których zakres znaczeniowy jest zbliżony lub identyczny. zamiennie zamiast określenia "klasyfikacja" używa się takich terminów jak: grupowanie, podział, dyskryminacja, delimitacja, taksonomia, analiza typu Q, analiza skupień, identyfikacja. Ta różnorodność terminologiczna wynika między innymi z faktu, że metody klasyfikacji tworzone i stosowane są przez przedstawicieli różnych dyscyplin badawczych /biologowie, antropologowie, psychologowie, ekonomiści, matematycy/. Trudność usystematyzowania tych pojęć jest wynikiem tego, że te same terminy są rozmaicie definiowane przez różnych autorów /najczęściej twórców lub współtwórców odpowiednich dyscyplin/ [68]. Nie wnikając szczegółowo w zakres pojęciowy wymienionych terminów w niniejszej pracy - zgodnie z jej tytułem - używany będzie termin klasyfikacja.

T. Wójcik w pracy [115] wyróżnia następujące koncepcje klasyfikacji:

- 1^o językowa koncepcja klasyfikacji,
- 2^o pozajęzykowa koncepcja klasyfikacji.

Koncepcja pierwsza reprezentowana jest w literaturze przez takich teoretyków jak K. Ajdukiewicz i T. Kotarbiński. Według tej koncepcji klasyfikacją jest pewien odpowiednio skonstruowany zespół zwrotów językowych - zespół nazw. Można tę koncepcję bardziej szczegółowo wyrazić w następujący sposób [115, s.17] :

a/ klasyfikowaniu podlega zbiór przedmiotów /zakres nazwy/;

b/ klasyfikacja jest zespołem nazw, między zakresami których zachodzą określone stosunki;

c/ klasyfikowanie jest czynnością tworzenia klasyfikacji /zespołu nazw/ dla klasyfikowanego zbioru przedmiotów /zakresu nazwy/.

"Dobra" klasyfikacja powinna być zdaniem T. Kotarbińskiego [115, s.16] wyczerpująca i rozłączna. Wyczerpująca jest wtedy, gdy suma zakresów, na które podzielono zakres danej nazwy, równa się całemu temu zakresowi a rozłączna polega na wyłączaniu się wzajemnym owych zakresów cząstkowych.

Językowa koncepcja klasyfikacji wykorzystana została w głównej mierze w podrozdziale 1.3 poświęconemu terminologicznej klasyfikacji struktur. Koncepcja ta wykorzystywana jest również w innych rozdziałach pracy np. w trzecim, gdzie miary podobieństwa dzieli się na miary odległości i bliskości czy szóstym, gdzie przemysł podzielono na grupy gałęzi przemysłu.

W najogólniejszym sensie pozajęzykową koncepcję klasyfikacji definiuje się w sposób następujący /por. [9] /:

Określenie 4.1 Klasyfikacja jest to podział zbioru obiektów na klasy.

Klasyfikacją w podanym sensie jest więc zarówno zupełnie dowolny /bezładny/ podział zbioru obiektów, jak i podział uwzględniający określone zalecenia formalne, zawężające w istocie rzeczy zakres pojęcia "klasyfikacja" [9]. Taka definicja dla potrzeb niniejszej pracy ma zbyt szeroki zakres pojęciowy, dlatego też poniżej przedstawione zostanie szczegółowe określenie klasyfikacji, w którym zawęży się zakres pojęcia klasyfikacji przez podanie dodatkowych warunków formalnych.

Określenie 4.2 Klasyfikacja jest to podział /aglomeracyjny lub deaglomeracyjny/ zupełny niejednorodnego - z punktu widzenia diagnostycznych cech strukturalnych - zbio-

ru obiektów na nieustaloną a priori liczbę względnie jednorodnych klas /niepustych i rozłącznych/ na podstawie relacji podobieństwa obiektów.

Klasyfikację zbioru obiektów przeprowadza się w oparciu o unormowane obrazy struktur tych obiektów /por. rozdział drugi/. Klasyfikacja przez podział aglomeracyjny rozpoczyna się od sytuacji, w której każdy obiekt badania tworzy początkowo odrębną klasę, natomiast podział deglomeracyjny rozpoczyna się od jednej klasy obejmującej wszystkie obiekty.

Zatem tak postawione zagadnienie klasyfikacji zbioru A , zawierającego obiekty A_1, A_2, \dots, A_n , na klasy P_1, P_2, \dots, P_H musi spełniać następujące warunki:

1° zupełności

$$\bigcup_{h=1}^H P_h = A, \quad /4.1/$$

2° rozłączności

$$\begin{array}{c} \wedge \\ h, h' = 1, 2, \dots, H \\ h \neq h' \end{array} P_h \cap P_{h'} \neq \emptyset, \quad /4.2/$$

3° niepustości

$$\bigwedge_h P_h \neq \emptyset. \quad /4.3/$$

Określenie klasyfikacji według koncepcji pozajęzykowej zawiera w sobie niezdefiniowany termin klasa jednorodna /homogeniczna/, którego zakres znaczeniowy jest zbliżony do następujących wyrazów: grupa, typ, skupienie, takson. W literaturze różnie definiuje się to pojęcie [19; 21] :

1° klasa jest taką zbiorowością obiektów w przestrzeni, w której podobieństwo pomiędzy dowolną parą obiektów jest większe niż podobieństwo

bieństwo pomiędzy jakimkolwiek obiektem należącym do klasy, a dowolnym obiektem nie należącym do niej;

2^o klasy tworzą obiekty, których podobieństwo z najbardziej im podobnym obiektem jest większe niż podobieństwo najbardziej podobnych obiektów należących do różnych klas;

3^o klasy budowane na zasadzie gwiazdy oraz klasy pełne. Gwiazda jest to zbiór obiektów, których podobieństwo do ustalonego obiektu - zwanego środkiem gwiazdy - jest nie mniejsze niż wartość progowa φ . Klasa pełna jest to zbiór obiektów, w którym podobieństwo każdej pary obiektów jest nie mniejsze niż ustalona wartość progowa φ ;

4^o klasy tworzą obiekty jak najbardziej podobne, natomiast w różnych klasach znajdują się obiekty jak najmniej podobne;

5^o klasami są takie obszary w przestrzeni, które charakteryzują się większą gęstością i są oddzielone obszarami o mniejszej gęstości;

6^o klasy tworzą obiekty, których podobieństwo z najbardziej im podobnym obiektem jest większe niż wszystkie podobieństwa międzyklasowe /te z kolei mogą być zdefiniowane w różny sposób - patrz tab. 4.1/.

Definicja 4^o - z uwagi na jej ogólny charakter /definicje 1^o - 3^o oraz 6^o są jej szczególnymi przypadkami/ - ma w praktyce największe znaczenie, bowiem dopuszcza stosowanie różnych rozwiązań w zakresie metod klasyfikacji.

Wymienione wyżej definicje klasy nie wyczerpują istniejących. Inne propozycje podane są między innymi w cytowanej już pracy B. S. Everitta oraz w pracy C. R. Rao [84] .

Po tym krótkim omówieniu podstawowych pojęć związanych z klasyfikacją przedstawione zostaną najistotniejsze własności metod klasyfikacji. Znajomość określonych własności metod klasyfikacji pozwala na właściwe ich wykorzystanie w badaniach empirycznych [80].

Pierwsze własności formalne wypracowali N. Jardine i R. Sibson [44; 45]. Zostały one następnie wzbogacone nowymi w pracach [25; 104], a zwięzły ich przegląd w literaturze zawierają następujące pozycje [2; 41; 80].

Wybrane własności formalne metod klasyfikacji można ująć w postaci następujących punktów:

- 1^o własność jednoznaczności /najlepszego obrazu/; jest to warunek konieczny, który powinna spełniać każda metoda klasyfikacji; własność ta oznacza, że kolejność w jakiej obiekty są klasyfikowane nie ma wpływu na wynik klasyfikacji; inaczej mówiąc podziału na klasy nie można poprawić przez zmianę wstępnego uporządkowania obiektów;
- 2^o własność poprawnej struktury według drzewka połączeń; /dla hierarchicznych metod aglomeracyjnych/; metody klasyfikacji posiadają tę własność, jeżeli rezultaty klasyfikacji hierarchicznej zbioru obiektów otrzymane w wyniku ich zastosowania dadzą się przedstawić w postaci drzewka połączeń /dendrogramu/ zgodnego z kolejnością podobieństwa pomiędzy elementami tego zbioru /wartości poziomu połączenia klas rozłożone są monotonicznie rosnąco w przypadku miar odległości oraz monotonicznie malejąco w przypadku miar bliskości/; definicja dendrogramu podana zostanie w podrozdziale 4.2.1;
- 3^o własność poprawnej struktury według klas; klasy tworzą obiekty jak najbardziej podobne, natomiast w różnych klasach są obiekty jak najmniej podobne; warunek ten dopuszcza stosowanie różnych rozwiązań w zakresie metod klasyfikacji /związany jest on z pojęciem klasy/; zakłada się, że metoda klasyfikacji posiada własność poprawnej struktury według klas, gdy w jej wyniku otrzymuje się klasy obiektów, których podobieństwo z najbardziej im podobnym obiektem jest większe niż wszystkie podobieństwa międzyklasowe /patrz szósta definicja klasy/;

4^o własność wypukłości: metody klasyfikacji posiadają tę własność, jeżeli w rezultacie ich zastosowania otrzymuje się podział zbioru obiektów $P = \{P_1, P_2, \dots, P_H\}$, w którym wypukłe otoczenia klas P_1, P_2, \dots, P_H nie przecinają się [2].

Oprócz własności formalnych istnieją również opracowane przez szkołę australijską pewne zalecenia praktyczne ważne z punktu widzenia potencjalnego użytkownika metod klasyfikacji [114]:

1^o klasyfikacja musi być bardziej intensywna niż to implikują stosowane miary podobieństwa; oznacza to, że w wyniku klasyfikacji powinno otrzymać się mniejszą liczbę wyraźnie określonych klas; taki rezultat jest pożyteczny wtedy, gdy użytkownika bardziej interesują granice pomiędzy klasami, aniżeli struktura wewnętrzna klas /np. w ekologii/;

2^o klasyfikacja musi być relatywnie niewrażliwa na wartości odosobnione /krańcowe/; na wynik klasyfikacji nie powinny mieć silnego wpływu obiekty odosobnione /znajdujące się w oddali od innych obiektów zbioru A/ oraz błędy i nieścisłości w danych.

W literaturze spotkać można wiele propozycji językowej klasyfikacji metod klasyfikacji [4; 12; 17; 21; 42; 81; 89]. Nie istnieje jednak rozłączny podział tychże metod uwzględniający choćby większość propozycji.

W pracy zaprezentowany zostanie dwustopniowy podział metod klasyfikacji. W pierwszym stopniu dzieli się je na hierarchiczne i niehierarchiczne, a te z kolei w stopniu drugim na aglomeracyjne i deglomeracyjne. Taki podział będzie w pełni wystarczający, jako że w niniejszej pracy ograniczono się do przeglądu metod klasyfikacji opartych na macierzy podobieństw.

Zasadnicza różnica między klasyfikacją hierarchiczną i niehierarchiczną polega na tym, że w wyniku przeprowadzenia pierwszej z nich otrzymuje się ciąg podziałów /przy czym liczba tych podziałów może zawierać się w przedziale $\langle 2; N/A - 1 \rangle$, zakładając że

nie wliczamy podziału wyjściowego/, natomiast przeprowadzając drugą otrzymuje się jeden podział wynikowy. Dodatkowo w klasyfikacji hierarchicznej musi być spełniony taki warunek, że podział w danym kroku klasyfikacji powstaje z połączenia /rozpadu/ dwóch /jednej/ lub więcej klas w kroku poprzednim w przypadku klasyfikacji aglomeracyjnej /deglomeracyjnej/.

Większość metod klasyfikacji ma uniwersalny charakter, tzn. niezależna jest od przyjętej miary podobieństwa obiektów. W pracy nie zaprezentowano /por. podrozdział 4.2/ żadnych metod, które służyłyby tylko do klasyfikacji obiektów ze względu na obrazy ich struktur. Wynika to z oczywistego faktu, że istota podejścia strukturalnego na etapie metod klasyfikacji nie różni się od podejścia w zwykłym sensie. Odmienność podejścia strukturalnego od innych typów badań została ujęta m. in. w rozdziale drugim i trzecim. W podrozdziale 3.1 wprowadzono dodatkowy obligatoryjny warunek dla miar podobieństwa struktur obiektów, który wyeliminował z badań strukturalnych te metody, które opierają się na macierzy podobieństw policzonej w oparciu o miarę odległości Euklidesowej /bądź jej kwadrat/. Omówienie tych metod klasyfikacji zawiera wiele monografii [4; 17; 21; 37] .

W związku z tym, że jest bardzo dużo metod klasyfikacji /wybrane omówione zostaną w podrozdziale 4.2/ przed ich użytkownikiem stoi nie lada problem, tzn. którą /bądź które/ z nich rekomendować do badań strukturalnych. Trzeba sobie w tym miejscu zdać sprawę z tego, że na pytanie która z metod klasyfikacji jest optymalna w badaniach strukturalnych nigdy nie uzyskamy jednoznacznej odpowiedzi. Analizując metody klasyfikacji od strony ich wad i zalet praktycznych oraz znając ich pewne własności formalne można zawęzić nieco liczbę metod, które powinny stosować się w danym badaniu. Z kolei wybór jednej z nich będzie wyborem arbitralnym.

W niektórych procedurach metod klasyfikacji zostały przeprowadzone pewne modyfikacje /nie zmieniające istoty tych metod/ celem uczynienia ich bardziej zrozumiałymi.

4.2. Przegląd wybranych metod klasyfikacji.

4.2.1. Metody hierarchiczne.

Wśród metod hierarchicznych wyróżnia się metody aglomeracyjne oraz deglomeracyjne. Punktem wyjścia tych metod jest symetryczna macierz podobieństw o postaci /3.26/ zaprezentowana w rozdziale trzecim.

M E T O D Y A G L O M E R A C Y J N E

Klasyfikacja aglomeracyjna zwana także klasyfikacją indukcyjną, wstępującą lub też klasyfikacją przez grupowanie rozpoczyna się od sytuacji, w której każdy obiekt badania A_r / $r = 1, 2, \dots, n$ / tworzy początkowo jedną klasę P_h / $h = 1, 2, \dots, H$; gdzie $H = n$ /. W związku z tym macierz podobieństw $[C]$ przyjmuje następującą postać:

$$[C] = \begin{bmatrix} c/P_1, P_1/ & c/P_1, P_2/ & \dots & c/P_1, P_n/ \\ c/P_2, P_1/ & c/P_2, P_2/ & \dots & c/P_2, P_n/ \\ \vdots & \vdots & & \vdots \\ c/P_n, P_1/ & c/P_n, P_2/ & \dots & c/P_n, P_n/ \end{bmatrix} \quad /4.4/$$

Hierarchiczne metody aglomeracyjne charakteryzują się /w ujęciu klasycznym¹/ następującymi cechami:

1. punktem wyjścia jest N/A klas jednoelementowych /jest tyle klas ile jest obiektów/; N/A - liczba obiektów w zbiorze A ;

¹W ujęciu nieklasycznym proces klasyfikacji można zakończyć w ciągu mniejszej liczby kroków niż $N/A - 1$ /por. [12] /.

2. po każdym kroku klasyfikacji liczba klas zmniejsza się o jeden, przy czym zmniejszenie liczby klas następuje przez połączenie dwóch istniejących;
3. istnieje $N/A/ - 1$ kroków klasyfikacji; po $N/A/ - 1$ krokach otrzymuje się jedną klasę zawierającą wszystkie obiekty $A_r / r = 1, 2, \dots, n/$;
4. proces klasyfikacji można przedstawić graficznie przy pomocy dendrogramu /drzewka połączeń/, wskazującego na kolejność połączeń między klasami.

N. Jardine i R. Sibson [44, s.49] definiują dendrogram w sposób następujący. Niech $E/A/$ oznacza zbiór relacji równoważności /ekwiwalentności/ określony na zbiorze A . Dendrogram jest to funkcja $f : [0; 1] \rightarrow E/A/$ spełniająca następujące warunki:

a/ każda klasa na danym poziomie \mathcal{L}' jest połączeniem klas na poziomie \mathcal{L} , gdzie

$$0 \leq \mathcal{L} \leq \mathcal{L}' \text{ /dla miar odległości/},$$

$$1 \geq \mathcal{L} \geq \mathcal{L}' \text{ /dla miar bliskości/};$$

Poziom połączenia klas jest to wartość podobieństwa między łączonymi klasami w danej iteracji hierarchicznej metody aglomeracyjnej /wartości te dla poszczególnych metod odczytuje się w kroku 1 centralnej procedury aglomeracyjnej - patrz tab. 4.1 i 4.2/.

b/ dla dostatecznie dużej /małej - dla miar bliskości/ wartości \mathcal{L} wszystkie obiekty znajdują się w jednej klasie;

c/ mając daną wartość \mathcal{L} , istnieje takie $\delta > 0 / \delta < 0$ - dla miar bliskości/, że klasy na poziomie \mathcal{L} i $\mathcal{L} + \delta$ są identyczne.

Ta definicja dendrogramu odrzuca zatem te hierarchiczne metody klasyfikacji, które nie posiadają własności poprawnej struktury według drzewka połączeń.

Hierarchiczne metody aglomeracyjne działają według centralnej procedury aglomeracyjnej [4]. Algorytm tej procedury jest następujący:

1. Szuka się w macierzy podobieństw pary klas najbardziej podobnych. Niech otrzymanymi będą klasy o numerach t i q / $t > q$ /.
2. Redukuje się liczbę klas o jeden łącząc klasy o numerach t i q w nową, nadając jej numer q i usuwając klasę o numerze t .
3. Przekształca się podobieństwa /stosownie do metody/ pomiędzy klasą o numerze q i pozostałymi klasami.
4. Powtarza się kroki 1 - 3 do chwili, gdy wszystkie obiekty A_r / $r = 1, 2, \dots, n$ / znajdują się w jednej klasie.

Różnice w procedurach metod aglomeracyjnych wynikają ze sposobu definiowania klas najbardziej podobnych w kroku pierwszym i z odmienności pojmowania podobieństwa klas w etapie trzecim.

W zależności od tego czy metody aglomeracyjne oparte są na macierzy odległości czy blikości to różnice w kroku pierwszym i trzecim centralnej procedury aglomeracyjnej można przedstawić odpowiednio w tabeli 4.1 i 4.2 [4; 5; 17; 21; 50; 89; 98].

W tym miejscu przedstawiona zostanie krótka charakterystyka metod aglomeracyjnych.

Nazwa metody pojedynczego połączenia wywodzi się stąd, że klasy w niej łączy się na zasadzie największego podobieństwa spośród wszystkich podobieństw między obiektami należącymi do łączonych klas. Podstawową cechą tej metody jest skłonność do tworzenia klas w kształcie serpentyny [21]. Efekt ten czasami zwany jest łańcuchem /od angielskiego słowa chaining/. Skutkiem skłonności do tworzenia łańcuchów jest możliwość powstawania klas zawierających obiekty mało do siebie podobne. Stosowanie tej metody prowadzi do dołączania obiektów do klas już istniejących, natomiast rzadziej powstają nowe klasy. Zjawisko to może być pożyteczne w przypadku, gdy nie interesuje nas poszukiwanie klas homoge-

Tabela 4.1

Metody klasyfikacji oparte na macierzy odległości

Lp.	Nazwa metody	Krok 1	Krok 3
1	2	3	4
1	pojedynczego połączenia	$d/P_t, P_q / = \min d/P_r, P_s /$	$d/P_q, P_u / = \min \{ d/P_q, P_u /; d/P_t, P_u / \}$
2	kompletnego połączenia	$d/P_t, P_q / = \min d/P_r, P_s /$	$d/P_q, P_u / = \max \{ d/P_q, P_u /; d/P_t, P_u / \}$
3	średniego połączenia wewnątrz nowej klasy	$\frac{\text{sum } P_t + \text{sum } P_q + d/P_t, P_q /}{\left[\frac{N/P_t / + N/P_q /}{\text{sum } P_r + \text{sum } P_s + d/P_r, P_s /} + 1 \right] / 2} =$ $= \min \frac{\left[\frac{N/P_r / + N/P_s /}{\text{sum } P_r + \text{sum } P_s + d/P_r, P_s /} + 1 \right] / 2}{\left[\frac{N/P_r / + N/P_s /}{\text{sum } P_r + \text{sum } P_s + d/P_r, P_s /} + 1 \right] / 2}$	$d/P_q, P_u / = d/P_q, P_u / + d/P_t, P_u /$
4	średniego połączenia międzyklasowego	$\frac{d/P_t, P_q /}{N/P_t / N/P_q /} = \min \frac{d/P_r, P_s /}{N/P_r / N/P_s /}$	$d/P_q, P_u / = d/P_q, P_u / + d/P_t, P_u /$
4.1	Sposób A	$d/P_t, P_q / = \min d/P_r, P_s /$	$d/P_q, P_u / = \frac{N/P_q /}{N/P_q / + N/P_t /} d/P_q, P_u / +$ $+ \frac{N/P_t /}{N/P_q / + N/P_t /} d/P_t, P_u /$
4.2	Sposób B		

1	2	3	4
5	skupiania parami	$d/P_t, P_q / = \min d/P_r, P_s /$	$d/P_q, P_u / = \frac{1}{2} \left\{ d/P_q, P_u / + d/P_t, P_u / \right\}$
6	"środek ciężkości"	$d/P_t, P_q / = \min d/P_r, P_s /$	$d/P_q, P_u / = \frac{N/P_q /}{N/P_t / + N/P_q /} d/P_q, P_u / +$ $+ \frac{N/P_t /}{N/P_t / + N/P_q /} d/P_t, P_u / - \frac{N/P_t / N/P_q /}{[N/P_t / + N/P_q /]^2}$ $d/P_q, P_t /$
7	"mediany"	$d/P_t, P_q / = \min d/P_r, P_s /$	$d/P_q, P_u / = \frac{1}{2} \left\{ d/P_q, P_u / + d/P_t, P_u / \right\} -$ $\frac{1}{4} d/P_q, P_t /$
8 8.1	Metody Farris'a wersja A /wagi różnicowane/	$d/P_t, P_q / = d/P_r, P_s /$, przy czym dla klas o numerach r i s funkcja $\frac{N/P_r / N/P_s /}{N/P_r / + N/P_s /} \sum_{\substack{u=1 \\ u \neq r, s}}^z \frac{N/P_u /}{N/P_u /} \left[d/P_r, P_u / - \right.$ $\left. - d/P_s, P_u / \right]^2$ osiąga minimum	$d/P_q, P_u / = \frac{N/P_q / d/P_q, P_u / + N/P_t / d/P_t, P_u /}{N/P_q / + N/P_t /}$

1	2	3	4
8.2	wersja B /wagi jedna- kowe/	<p> $d/P_t, P_q = d/P_r, P_s$, przy czym dla klas o numerach r i s funkcja </p> $\frac{1}{2} \sum_{u=1}^2 \left[d/P_r, P_u - d/P_s, P_u \right]^2$ <p> $u \neq r, s$ </p> <p>osiąga minimum</p>	$d/P_q, P_u = \frac{1}{2} \left[d/P_q, P_u + d/P_t, P_u \right]$

Metody klasyfikacji oparte na macierzy bliskości

Lp.	Nazwa metody	Krok 1	Krok 3
1	pojedynczego połączenia	$\epsilon/P_t, P_q/ = \max \epsilon/P_r, P_s/$	$\epsilon/P_q, P_u/ = \max \{ \epsilon/P_q, P_u/; \epsilon/P_t, P_u/ \}$
2	kompletnego połączenia	$\epsilon/P_t, P_q/ = \max \epsilon/P_r, P_s/$	$\epsilon/P_q, P_u/ = \min \{ \epsilon/P_q, P_u/; \epsilon/P_t, P_u/ \}$
3	średniego połączenia wewnątrz nowej klasy	$\frac{\text{sum } P_t + \text{sum } P_q + \epsilon/P_t, P_q/}{[N/P_t/ + N/P_q/][N/P_t/ + N/P_q/ - 1]/2} =$ $= \max \frac{\text{sum } P_r + \text{sum } P_s + \epsilon/P_r, P_s/}{[N/P_r/ + N/P_s/][N/P_r/ + N/P_s/ - 1]/2}$	$\epsilon/P_q, P_u/ = \epsilon/P_q, P_u/ + \epsilon/P_t, P_u/$
4	średniego połączenia międzyklasowego	$\frac{\epsilon/P_t, P_q/}{N/P_t/ + N/P_q/} = \max \frac{\epsilon/P_r, P_s/}{N/P_r/ + N/P_s/}$	$\epsilon/P_q, P_u/ = \epsilon/P_q, P_u/ + \epsilon/P_t, P_u/$
4.1	Sposób A	$\epsilon/P_t, P_q/ = \max \epsilon/P_r, P_s/$	$\epsilon/P_q, P_u/ = \frac{N/P_u/}{N/P_q/ + N/P_t/} \epsilon/P_q, P_u/ + \frac{N/P_t/}{N/P_t/ + N/P_q/} \epsilon/P_t, P_u/$
4.2	Sposób B	$\epsilon/P_t, P_q/ = \max \epsilon/P_r, P_s/$	$\epsilon/P_q, P_u/ = \frac{1}{2} \{ \epsilon/P_q, P_u/ + \epsilon/P_t, P_u/ \}$
5	Skupiania parami	$\epsilon/P_t, P_q/ = \max \epsilon/P_r, P_s/$	$\epsilon/P_q, P_u/ = \frac{1}{2} \{ \epsilon/P_q, P_u/ + \epsilon/P_t, P_u/ \}$

1	2	3	4
6	"Średnia ciężkości"	$C/P_t, P_q / = \max C/P_r, P_s /$	$C/P_q, P_u / = \frac{N/P_q /}{N/P_t / + N/P_q /} C/P_q, P_u / +$ $+ \frac{N/P_t /}{N/P_t / + N/P_q /} C/P_t, P_u / +$ $+ \frac{N/P_t / N/P_q /}{[N/P_t / + N/P_q /]^2} / 1 - C/P_q, P_t //$
7	"mediany"	$C/P_t, P_q / = \max C/P_r, P_s /$	$C/P_q, P_u / = \frac{1}{2} \{ C/P_q, P_u / + C/P_t, P_u / \} +$ $+ \frac{1}{4} / 1 - C/P_q, P_t /$

Legenda /dotyczy tab. 4.1 i 4.2/:

$N/P_u /$, $N/P_t /$, $N/P_q /$, $N/P_r /$, $N/P_s /$ - liczebności obiektów w klasie o numerze odpowiednio u, t, q, r i s;
 $\text{sum } P_t$, $\text{sum } P_q$, $\text{sum } P_r$, $\text{sum } P_s$ - oznaczają sumy wszystkich podobieństw między obiektami należącymi do klas t, q, r i s /odczytane z wyjściowej macierzy podobieństw/;
z - aktualna liczba klas.

Źródło: opracowanie własne na podstawie prac [4; 23; 50; 89].

nicznych, a jedynie klas o optymalnym połączeniu obiektów [21, s.61]. W praktyce jednak użytkownicy metod klasyfikacji dążą do uzyskania klas homogenicznych, przeto efekt łańcucha stanowi wadę tej metody. Skłonność do tworzenia klas w postaci łańcuchów cechuje również inne metody hierarchiczne /za wyjątkiem metody kompletnego połączenia/, z tym że w metodzie pojedynczego połączenia przejawia się ona najsilniej. Wynika to z faktu, że metody te nie posiadają własności wypukłości. W wyniku zastosowania metody pojedynczego połączenia otrzymuje się najkrótszą sieć połączeń między obiektami zbioru A.

Nazwa metody kompletnego połączenia pochodzi stąd, że klasy w niej łączy się na zasadzie najmniejszego podobieństwa spośród wszystkich podobieństw między obiektami należącymi do łączonych klas /kryterium to uwzględnia więc wszystkie połączenia między obiektami łączonych klas/. W efekcie zastosowania tej metody otrzymuje się najdłuższą sieć połączeń między obiektami zbioru A.

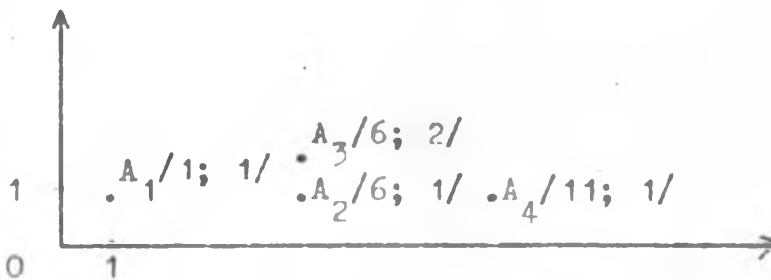
Wadą dwóch powyższych metod jest oparcie się przy ustalaniu podobieństwa klas na wartościach krańcowych. Wszystkie pozostałe metody przyjmują wartości pośrednie przy ustalaniu podobieństwa klas. Wśród tych metod wyróżnia się takie, które uwzględniają liczbę obiektów w klasie przy ustalaniu podobieństwa międzyklasowego /metoda średniego połączenia wewnątrz nowej klasy, metoda średniego połączenia międzyklasowego, metoda "środką ciężkości"/ jak i metody, które łączone klasy traktują równoważnie, bez względu na liczbę obiektów w klasie /metoda "mediany", metoda skupiania parami/.

Nazwy dwóch metod, a mianowicie "środką ciężkości" i "mediany" zostały zapisane w cudzysłowie, ponieważ posiadają one interpretację geometryczną tylko dla kwadratu odległości Euklidesowej. Miara ta w niniejszej pracy nie jest stosowana, jako że nie spełnia wszystkich warunków ustalonych w rozdziale trzecim dla miar podobieństwa struktur. Metody te mogą być stosowane /por. [4, s.141]/,

gdy macierz podobieństw jest liczona w oparciu o inne miary odległości lub bliskości, ale interpretacja tak otrzymanych wyników /w sensie podobieństwa międzyklasowego/ nie jest zgodna z nazwami tych metod /przekształcenie w kroku 3 tych metod nie posiada wartości interpretacyjnej/.

Charakterystyczną cechą metod centroidalnych /"środką ciężkości" i "mediany"/ oraz średniego połączenia wewnątrz nowej klasy jest to, że wartości poziomu podobieństwa \mathcal{L} związane z łączeniem najbardziej podobnych klas mogą podnosić się i spadać przy przechodzeniu z kroku na krok w klasyfikacji hierarchicznej /por. [4, s.141]/, czyli nie posiadają te metody własności poprawnej struktury według drzewka połączeń.

J. S. Farris w pracy [23] zaproponował dwie metody klasyfikacji działające według centralnej procedury aglomeracyjnej. Krok 1 i 3 w tych metodach zaprezentowano w tabeli 4.1 /pozycja ósma/. Na prostym przykładzie można wykazać, że przyjęte w tych metodach kryterium łączenia klas najbardziej podobnych w kroku 1 jest błędne. Niech dane będą następujące obserwacje na obiektach A_r / $r = 1, 2, 3, 4$ / w przestrzeni dwuwymiarowej:



Wykorzystując do pomiaru podobieństwa obiektów odległość miejską /por. np. [79, s.15] / wylicza się symetryczną macierz podobieństw:

	1	2	3	4
1	0	5	6	10
2	5	0	1	5
3	6	1	0	6
4	10	5	6	0

W iteracji pierwszej łączy się w jedną klasę o numerach 1 i 4. Dla tych klas /zarówno w wersji A jak i B/ funkcja podana w tabeli 4.1 osiąga wartość minimalną równą zero. W pierwszej kolejności połączono więc za pomocą tych metod dwie klasy, które są najbardziej niepodobne. Już po iteracji pierwszej można powiedzieć, że nie spełniona jest żadna z definicji klasy podanych w podrozdziale 4.1.

M E T O D Y D E G L O M E R A C Y J N E

Klasyfikacja deglomeracyjna zwana jest także klasyfikacją dedukcyjną, zstępującą lub klasyfikacją przez podział.

Hierarchiczne metody deglomeracyjne charakteryzują się następującymi cechami:

1. punktem wyjścia jest jedna klasa obejmująca wszystkie obiekty badania A_1, A_2, \dots, A_n ;
2. w każdym kroku klasyfikacji liczba klas zwiększa się o jeden, przy czym zwiększenie liczby klas następuje przez rozdzielenie jednej z istniejących klas;
3. istnieje $N/A/ - 1$ kroków klasyfikacji; po $N/A/ - 1$ krokach otrzymuje się liczbę klas równą liczbie obiektów badania $N/A/$, tzn. każdy obiekt tworzy jedną klasę.

Poniżej zaprezentowane zostaną trzy metody Huberta [50; 98] działające według ogólnej procedury deglomeracyjnej oraz metoda taksonomii wrocławskiej [27; 34; 72] i najkrótsza sieć połączeń Prima [34; 81].

Algorytm postępowania w metodach Huberta działających według ogólnej procedury deglomeracyjnej jest następujący:

1. Dla każdej istniejącej klasy wyznaczyć parę obiektów najbardziej niepodobnych i spośród tych par wybrać tę, dla której podobieństwo to jest najmniejsze /w pierwszej iteracji będzie tylko jedna para/. Niech będą to obiekty A_t i A_q należące do klasy P_h .

2. Klasa P_h zostaje podzielona na dwie klasy o numerach t i q . Przydzielić obiekt A_t do klasy P_t , a obiekt A_q do klasy P_q . Usunąć obydwa obiekty z klasy P_h .

3. Rozdzielić pozostałe obiekty klasy P_h między klasy P_t i P_q . Sposób podziału jest różny dla każdej z wersji.

Wersja A. Dla każdego obiektu pozostałego w klasie P_h wyznaczyć najmniej podobny obiekt z klas P_t i P_q i wybrać ten z klasy P_h dla którego podobieństwo to jest najmniejsze. Jeśli jest on najmniej podobny do obiektu będącego w klasie P_t /odpowiednio P_q / przydzielić go do klasy P_q /odpowiednio P_t / i usunąć z klasy P_h .

Wersja B. Dla każdego obiektu pozostałego w klasie P_h wyznaczyć najbardziej podobny obiekt z klas P_t i P_q i wybrać ten z klasy P_h dla którego podobieństwo to jest największe. Przydzielić go do tej klasy, w której znajduje się obiekt najbardziej do niego podobny i usunąć z klasy P_h .

Wersja C. Dla każdego obiektu pozostałego w klasie P_h wyznaczyć najbardziej podobny obiekt z klas P_t i P_q i wybrać ten z klasy P_h dla którego podobieństwo to jest najmniejsze. Przydzielić go do tej samej klasy, do której należy obiekt realizujący to podobieństwo. Usunąć wyznaczony obiekt z klasy P_h .

Krok 3 powtarzany jest do chwili wyczerpania obiektów w klasie P_h .

4. Z listy istniejących klas usunąć klasę P_h i włączyć do niej klasy P_q i P_t . W ten sposób liczba klas została zwiększona o jeden.

Wadą hierarchicznych metod deglomeracyjnych Huberta jest to, że w każdym kroku podział determinowany jest najmniejszym podobieństwem wewnątrzklasowym dwóch obiektów, natomiast podobieństwa pozostałe nie są uwzględniane. Można tę niedogodność wyeliminować stosując w kroku 1 tej procedury inne kryterium uwzględniające wszystkie podobieństwa między obiektami w klasie. Autor niniejszej

pracy proponuje do tego celu wykorzystać średnie podobieństwo wewnątrzklasowe. Zatem treść kroku 1 jest następująca [106] :

1. Dla każdej istniejącej klasy policzyć wartość następującej funkcji:

$$\frac{\sum_{r=2}^{n_h} \sum_{s=1}^{s=r-1} c/A_r; A_s/}{\binom{n_h}{2}} \quad , \quad /4.5/$$

gdzie: n_h - liczba obiektów w klasie o numerze h ;

$$c/A_r; A_s/ = c[S/A_r/; S/A_s/] = c/P_{..r}; P_{..s}/ = c_{rs}.$$

Spośród tych klas wybiera się do podziału tę, dla której funkcja powyższa osiąga maksimum /minimum dla miar bliskości/. Dla klasy tej wyznacza się parę obiektów najbardziej niepodobnych. Niech będą to obiekty A_t i A_q należące do klasy P_h .

Pozostałe kroki nie ulegają zmianie.

Zaletą tego podejścia jest to, że w pierwszej kolejności zostają rozdzielone klasy o mniejszej zwartości przestrzennej. W przypadku kryterium wprowadzonego przez Huberta zwartość przestrzenna obiektów nie miała znaczenia, ponieważ o podziale klasy decydowało najmniejsze podobieństwo wewnątrzklasowe dwóch obiektów.

Metoda taksonomii wrocławskiej została opracowana na początku lat pięćdziesiątych przez pracowników naukowych Ogólnej Grupy Zastosowań Państwowego Instytutu Matematycznego we Wrocławiu, a opis jej podali autorzy w pracy [27] .

Procedura postępowania w tej metodzie jest następująca [34] :

1. W każdym wierszu /kolumnie/ macierzy podobieństw $[C]$ wyszukuje się elementu, który wskazuje parę obiektów najbardziej podobnych. Otrzymane połączenia między obiektami porządkuje się i przedstawia

w postaci grafu niezorientowanego I-go stopnia /wierzchołki tego grafu, czyli obiekty połączone są bezkierunkowo liniami/.

2. Sprawdza się czy graf niezorientowany jest dendrytem. Dendryt jest to łamana, która może się rozgałęziać, lecz nie może zawierać zamkniętych i taka, że każde dwa obiekty A_r i A_s zbioru A są przez nią połączone [27]. Jeśli okaże się, że otrzymano dendryt to przechodzimy do kroku 4, w przeciwnym przypadku do kroku 3.

3. Przechodzi się z grafu niezorientowanego rzędu β na graf rzędu $\beta + 1$ przez połączenie dwóch najbardziej podobnych klas w grafie rzędu β /klasy łączy się na zasadzie największego podobieństwa spośród wszystkich podobieństw między obiektami należącymi do łączonych klas/. Wraca się do kroku 2.

4. Ustawia się otrzymane w dendrycie podobieństwa między sąsiadującymi obiektami /graficznie reprezentowane są przez krawędzie/ w porządku wzrastającego podobieństwa. Odrzucając kolejno krawędzie łączące obiekty, zaczynając od krawędzi łączącej obiekty o najmniejszym podobieństwie, a kończąc na krawędzi łączącej obiekty o największym podobieństwie, otrzymuje się hierarchiczne uporządkowanie zbioru obiektów.

W trakcie budowy dendrytu metoda Prima [34; 81] zbiór badanych obiektów dzieli się na dwa podzbiory, z których pierwszy A_1 zawiera obiekty należące w danym momencie do tworzonego dendrytu, natomiast drugi A_2 - zawiera obiekty nie przyłączone jeszcze do dendrytu. Na początku zbiór A_1 jest zbiorem pustym, natomiast A_2 obejmuje wszystkie badane obiekty. W pierwszym kroku do zbioru A_1 włącza się dowolny obiekt należący do zbioru A_2 oraz korzystając z przyjętej miary podobieństwa oblicza się podobieństwa wybranego obiektu do pozostałych obiektów tworząc z nich zbiór B .

Dalsza procedura przebiega następująco:

1. Szuka się największego podobieństwa w zbiorze B , przyporządkowany mu obiekt przyłącza się do zbioru A_1 i jednocześnie wyklucza

ze zbioru A2.

2. Oblicza się podobieństwa wybranego obiektu do pozostałych obiektów zawartych w zbiorze A2, porównując je każdorazowo z odpowiadającymi im podobieństwami zawartymi w zbiorze B. Jeżeli podobieństwo któregoś obiektu okaże się większe niż dotychczasowe figurujące w zbiorze B, to dokonuje się ich zamiany, zostawiając w zbiorze B większe. W sytuacji odmiernej nie przeprowadza się zmian.

3. Sprawdza się czy zbiór A2 jest zbiorem pustym. Jeśli nie, to wraca się do kroku 1 i kontynuuje się powyższe operacje, aż w zbiorze A1 znajdą się wszystkie klasyfikowane obiekty.

4. Dendryt Prima dzieli się w analogiczny sposób jak dendryt wrocławski, otrzymując w efekcie hierarchiczną klasyfikację badanego zbioru obiektów.

Metody klasyfikacji budowane w oparciu o macierz podobieństw wymagają policzenia wartości $\frac{1}{2} N/A/ [N/A/ - 1]$ podobieństw /macierz podobieństw jest symetryczna, więc wystarczy policzyć tylko jeden trójkąt podobieństw w takiej macierzy/. W metodzie Prima liczenie podobieństw przebiega sekwencyjnie, tzn. zaczyna się w pierwszej iteracji od policzenia $N/A/ - 1$ podobieństw i kolejno $N/A/ - 2$, $N/A/ - 3$, ..., aż do 1 podobieństwa między obiektami, Otrzymuje się więc ciąg arytmetyczny. Należy wyliczyć sumę wyrazów tegoż ciągu, która wynosi:

$$\frac{N/A/ - 1 + 1}{2} [N/A/ - 1] = \frac{N/A/ [N/A/ - 1]}{2} . \quad /4.6/$$

Różnica między metodą Prima a metodą taksonomii wrocławskiej polega na tym, że:

a/ budowę dendrytu w metodzie Prima można rozpocząć w dowolnym miejscu /od dowolnego obiektu/;

b/ nie jest wymagana na początku algorytmu w metodzie Prima znajo-

mość całej macierzy podobieństw, a wyliczenie tej macierzy następuje sekwencyjnie dając w efekcie trójkąt podobieństw w symetrycznej macierzy podobieństw.

Wyniki otrzymane metodą Prima, taksonomii wrocławskiej oraz metodą pojedynczego połączenia są identyczne, jako że w rezultacie zastosowania każdej z nich otrzymuje się najkrótszą sieć połączeń. Metoda taksonomii wrocławskiej oraz metoda Prima charakteryzuje się takimi samymi cechami co metoda pojedynczego połączenia omówiona wcześniej.

4.2.2. Metody niehierarchiczne.

Ogół metod niehierarchicznych można podzielić na dwie grupy, tzn. na metody aglomeracyjne oraz deglomeracyjne. Punktem wyjścia poniżej omawianych metod będzie macierz podobieństw $[C]$ o postaci /3.26/ zaprezentowana w rozdziale trzecim.

M E T O D Y A G L O M E R A C Y J N E

Algorytm postępowania w niehierarchicznych metodach aglomeracyjnych jest następujący:

1. klasyfikacja rozpoczyna się od sytuacji, w której każdy obiekt badania A_r / $r = 1, 2, \dots, n$ / tworzy początkowo jedną klasę P_h / $h = 1, 2, \dots, H$; gdzie $H = n$; jest tyle klas ile jest obiektów, tzn. N/A ;
2. w jednym lub kilku krokach dokonuje się połączenia kilku klas w jedną; sposób połączenia zależy od stosowanej metody;
3. efektem końcowym jest klasyfikacja $P = \{P_1, P_2, \dots, P_H\}$ o liczbie klas N/P wynikającej z zastosowanej metody.

Istnieje bardzo dużo metod aglomeracyjnych realizujących po-

wyższy algorytm. Zaprezentowane zostaną w tym podrozdziale metody powszechnie stosowane jak i metody, które opierają się na diametralnie różnych kryteriach łączenia klas.

Prezentację rozpocznie bardzo prosta metoda progowa [42; 68]. Sposób postępowania w tej metodzie jest następujący. Na początku ustala się wielkość progową podobieństwa ϱ . W macierzy podobieństw $[C]$ w miejsce podobieństw między obiektami A_r i A_s mniejszych niż zadana wielkość progowa ϱ wstawia się zera, otrzymując macierz zredukowaną $[C_1]$. Na podstawie tej macierzy buduje się graf łączący tylko te obiekty, dla których podobieństwo w tej macierzy jest różne od zera. W ten sposób zostają połączone obiekty w klasy. Metoda ta jest bardzo pracochłonna dla dużych zbiorów obiektów.

Kolejne metody powstały w celu eliminacji efektu łańcucha występującego w metodzie pojedynczego połączenia. Wykorzystują one ten sam sposób definiowania podobieństwa międzyklasowego co w metodzie pojedynczego połączenia.

W trakcie budowy grafu w metodzie W. Pluty [75; 79] zbiór badanych obiektów A dzieli się na dwa podzbiory, z których pierwszy A_1 zawiera obiekty należące w danym momencie do utworzonych klas, natomiast drugi A_2 zawiera pozostałe obiekty. Na początku zbiór A_1 jest pusty, a A_2 obejmuje wszystkie badane obiekty. W pierwszym kroku wybiera się z macierzy podobieństw $[C]$ dwa obiekty o największym podobieństwie i włącza się je do zbioru A_1 /jednocześnie wykluczając z A_2 /, a podobieństwa tych obiektów do pozostałych zestawia się w zbiorze B_1 . Te dwa obiekty tworzą załączek pierwszej klasy.

Dalsza procedura przebiega następująco:

1. Tworzy się zbiór B_2 , w którym będą zapamiętywane po każdym kroku aktualne podobieństwa międzyklasowe /podobieństwem dwóch klas nazywa się największe podobieństwo między dwoma obiektami A_r i A_s , z których A_r należy do jednej klasy a A_s do drugiej/. Na początku w zbiorze B_2 figuruje wartość zero, jako że istnieje tylko jedna

klasa.

Szuka się największego podobieństwa w zbiorze B1, przyporządkowany mu obiekt przyłącza się do zbioru A1 i jednocześnie wyklucza ze zbioru A2. Wybrany obiekt może tworzyć klasę odosobnioną lub może wchodzić w skład jednej z istniejących klas, przy czym przyłączenie do klasy następuje wtedy, gdy spełniony jest warunek, że podobieństwa przyłączanego obiektu do obiektów danej klasy są większe od podobieństw danej klasy do klas pozostałych /patrz zbiór B2/.

2. Włącza się do zbioru B1 podobieństwa wybranego obiektu do wszystkich pozostałych w zbiorze A2, porównując je każdorazowo z odpowiadającymi podobieństwami zawartymi w zbiorze B1. Jeżeli podobieństwo któregoś obiektu okaże się większe niż dotychczasowe figurujące w zbiorze B1, to dokonuje się ich zamiany, zostawiając w zbiorze B1 podobieństwo większe. W przeciwnym przypadku nie przeprowadza się zmian.

3. Sprawdza się czy zbiór A2 jest pusty. Jeżeli nie, to wraca się do kroku 1 i kontynuuje się powyższe operacje, aż w zbiorze A1 znajdą się wszystkie klasyfikowane obiekty.

4. Ponieważ w wydzielonych klasach występuje zróżnicowanie pod względem średnic, dlatego też należy przeprowadzić ich ujednoczenie /średnicą klasy P_h nazywa się najmniejsze podobieństwo między dwoma obiektami należącymi do tej klasy/. Za podstawę ujednoczenia przyjmuje się średnicę S/P/ takiej klasy, dla której najmniejsze podobieństwo między dwoma obiektami w klasie jest najmniejsze spośród wszystkich wyodrębnionych wstępnie klas. Klasy, których podobieństwa międzyklasowe są większe niż S/P/ łączą się w jedną, jeśli tylko spełniony jest warunek, że najmniejsze podobieństwo między dwoma obiektami nowo tworzonej klasy /średnica klasy/ jest większe niż S/P/

Zaletą tej metody jest sprecyzowanie a priori kryterium łączenia klas /patrz krok 1/, które nie wymaga od potencjalnego użytkownika przyjmowania żadnych wartości krytycznych.

Zbliżony algorytm postępowania do metody grafowej W. Pluty posiada metoda TAXMAP, którą zaproponowali Carmichael i Sneath [21]. W metodzie TAXMAP nie jest realizowany 4 krok poprzednio omówionej procedury /kończy się na kroku 3/ oraz występuje różnica w kroku 1. Krok ten przebiega następująco:

1. Szuka się największego podobieństwa w zbiorze B₁, przyporządkowany mu obiekt przyłącza się do zbioru A₁ i jednocześnie wyklucza ze zbioru A₂. Warunkiem dołączenia wybranego obiektu do jednej z istniejących klas jest relatywnie niski spadek przeciętnego podobieństwa wewnątrzklasowego.

Problem w tej metodzie polega na ustaleniu co należy rozumieć przez relatywnie niski spadek przeciętnego podobieństwa wewnątrzklasowego. W pracy [21, s.31] podano wprawdzie pewien sposób postępowania, lecz wymaga on przyjęcia wartości progowej, która zdecyduje o tym czy obiekt będzie dołączony do klasy czy też nie. Należy to zaliczyć do słabości tej metody.

Inną bardzo prostą metodą pozwalającą eliminować w pewnym sensie efekt łańcucha w metodzie pojedynczego połączenia może być metoda opisana w pracy [2, s.102]. Algorytm jej zbliżony jest do metody W. Pluty z pominięciem 4 kroku, a krok 1 przebiega następująco:

1. Ustala się próg podobieństwa wewnątrzklasowego /jest to najmniejsze dopuszczalne podobieństwo między dwoma obiektami w klasie/. Szuka się największego podobieństwa w zbiorze B₁, przyporządkowany mu obiekt przyłącza się do zbioru A₁ i jednocześnie wyklucza ze zbioru A₂. Warunkiem dołączenia wybranego obiektu do jednej z klas jest nieprzekroczenie ustalonego progu podobieństwa.

Wadą tej metody jest potrzeba ustalenia wartości krytycznej oraz to, że ta wartość jest niezmienna dla wszystkich tworzonych klas.

W tym miejscu przedstawione zostaną dwie metody, w których klasy budowane są na zasadzie gwiazdy [19]. Gwiazdą nazywa się zbiór tych obiektów, których podobieństwo do ustalonego obiektu zwanego środkiem gwiazdy jest nie mniejsze niż wartość progowa ϱ /przy czym ϱ jest stałe lub zmienne przy wyodrębnianiu kolejnych klas/.

Algorytm postępowania w metodzie wrocławskiej [11] jest następujący:

1. Ustala się wartość krytyczną podobieństwa ϱ .
2. Rozpatruje się obiekty nie przydzielone jeszcze do klas. Dla każdego obiektu badania A_r / $r = 1, 2, \dots, n$ / znajduje się obiekty, których podobieństwo do r -tego jest nie mniejsze niż wartość krytyczna ϱ . Liczbę tych obiektów oznacza się symbolem N/A_r .
3. Spośród nie przydzielonych jeszcze do klas obiektów A_r wybiera się ten dla którego liczba obiektów N/A_r jest maksymalna. Jeżeli istnieje tylko jedno takie maksimum to przechodzi się do kroku 5, w przeciwnym przypadku do kroku 4.
4. Rozpatruje się sytuację w której istnieje więcej niż jeden obiekt o maksymalnej wartości N/A_r . Liczy się podobieństwa tych obiektów do obiektów, w których każda cecha ma współrzędne $\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k}$. Wybiera się ten obiekt dla którego podobieństwo to jest maksymalne.
5. Obiekt o maksymalnej wartości N/A_r wraz z obiektami, których podobieństwo do badanego jest większe niż wartość krytyczna ϱ tworzą klasę P_h . Rozpatruje się nowy zbiór, który powstaje przez usunięcie ze zbioru aktualnego obiektów należących do klasy P_h . Jeśli nowy zbiór nie jest zbiorem pustym to wraca się do kroku 2, w przeciwnym zaś przypadku kończy się postępowanie.

Metoda ta w decydującym stopniu uzależniona jest od wartości krytycznej ϱ /wartość ta jest niezmienna przy wyodrębnianiu kolejnych klas/. W wyniku zastosowania tej metody często otrzymuje się jedną klasę bardzo liczną i kilka mniej licznych.

Druga metoda wykorzystująca przy budowie klas zasadę gwiazdy została zaprezentowana w pracy [2, s.101]. Postępowanie w tej metodzie przebiega następująco:

1. Ustala się arbitralnie liczbę naturalną N / $1 \leq N \leq n$ /.
2. Dla każdego obiektu A_r / $r = 1, 2, \dots, n$ / znajduje się w macierzy $[C]$ podobieństwo do N -tego najbardziej podobnego obiektu. Otrzymuje się w ten sposób wartości krytyczne podobieństwa φ_r , które porządkuje się w kolejności od największego do najmniejszego podobieństwa i włącza do zbioru B_1 .
3. Spośród nie przydzielonych jeszcze do klas obiektów A_r wybiera się ten dla którego podobieństwo krytyczne φ_r / odczytane ze zbioru B_1 / jest największe. W przypadku, gdy istnieje kilka obiektów o równej największej wartości krytycznej podobieństwa φ_r to stosuje się postępowanie opisane w kroku 4 poprzedniej procedury.
4. Obiekt o największym podobieństwie φ_r wraz z obiektami, których podobieństwo do badanego jest większe niż wartość krytyczna φ_r tworzą klasę P_h . Rozpatruje się nowy zbiór obiektów, który powstaje przez usunięcie ze zbioru aktualnego obiektów należących do klasy P_h . Jeśli nowy zbiór nie jest zbiorem pustym to wraca się do kroku 3, w przeciwnym zaś przypadku kończy się postępowanie.

Cechą tej metody jest to, że pierwsza klasa otrzymana w wyniku jej algorytmu jest zawsze najliczniejsza i obejmuje $N + 1$ obiektów.

Kolejnym rodzajem klas, który wykorzystuje się w metodach klasyfikacji są klasy pełne. Klasa pełna jest to zbiór obiektów, w którym podobieństwo każdej pary obiektów jest nie mniejsze niż ustalona wartość progowa φ . W pracy [19] podano wprawdzie algorytm budowy klas pełnych, ale nie spełnia on podstawowego postulatu stawianego procedurom klasyfikacji, a mianowicie nie posiada własności najlepszego obrazu.

Poniżej podane zostaną dwie procedury klasyfikacji dające w

efekcie klasy pełne i spełniające kryterium najlepszego obrazu.

Algorytm metody eliminacji wektorów [13] przebiega następująco:

1. Ustala się krytyczną wartość progu podobieństwa ϱ .
2. Przekształca się macierz podobieństw $[C]$ w macierz $[C']$ w ten sposób, że w miejsce podobieństw między obiektami A_r i A_s nie mniejszych niż wartość krytyczna ϱ wpisuje się wartość 0, w przeciwnym zaś przypadku 1.
3. Oblicza się wektor kolumnowy $[C_0]$ sumując odpowiednie wiersze macierzy $[C']$.
4. Maksymalna wartość w wektorze $[C_0]$ wskazuje obiekt, który jest niepodobny na poziomie ϱ do największej liczby pozostałych obiektów. Wartość ta wskazuje, którą kolumnę i odpowiadający jej wiersz należy wyeliminować z macierzy $[C']$. Jeżeli w wektorze $[C_0]$ więcej niż jedna składowa jest równa wartości maksymalnej, to do wyboru wiersza i kolumny eliminowanych w danym kroku należy wykorzystać kryterium pomocnicze. Kryterium tym może być największa wartość podobieństwa między obiektami w danym wierszu lub największa suma podobieństw /wartości te odczytuje się z macierzy podobieństw $[C]$ /.
5. Czynności przedstawione w punktach 3 i 4 należy powtarzać do momentu, aż wszystkie składowe $[C_0]$ będą równe 0.
6. Obiekty odpowiadające nieskreślonym wierszom w macierzy $[C']$ tworzą klasę P_h .
7. Rozpatruje się nowy zbiór obiektów, który powstaje przez usunięcie ze zbioru aktualnego obiektów należących do wybranej klasy P_h . Jeśli nowy zbiór nie jest zbiorem pustym to wraca się do kroku 3 /wykreślając z macierzy zredukowanej $[C']$ wiersze i kolumny odpowiadające indeksom obiektów należących do wybranej klasy P_h /, w przeciwnym zaś przypadku kończy się postępowanie.

Rozwiązaniem problemu związanego z wyznaczaniem w każdym cyklu algorytmu wszystkich klas pełnych zajęła się autorka pracy [92]. Metoda ta posiada przeto walor optymalności. Algorytm jej jest

następujący:

1. Ustala się krytyczną wartość progu podobieństwa ϱ .
2. Przekształca się macierz podobieństw $[C]$ w macierz zredukowaną $[C']$ w ten sposób, że w miejsce podobieństw między obiektami A_r i A_s nie mniejszych niż wartość krytyczna ϱ wpisuje się wartość 0. Macierz ta pozwala na tworzenie klas pełnych P_h^a , a-elementowych $/a = 1, \dots, a_0; h = 1, 2, \dots, H/$.
3. Klasy 2-elementowe tworzy się przez wybór par obiektów $/A_r, A_s/$ dla których $c'/A_r, A_s/ = 0, r < s$.
4. Tworzy się teraz kolejno klasy 3-elementowe P_h^3 , tzn. dla każdej pary obiektów $/A_r, A_s/$ wybranej w kroku 3 dobiera się taki obiekt A_t , że $c'/A_r, A_t/ = 0$ oraz $c'/A_s, A_t/ = 0$.

Poszukiwanie klas a-elementowych należy zakończyć, jeżeli klasa $a_0 + 1$ elementowa jest pusta, tzn.:

$$P_h^{a_0+1} = \emptyset.$$

W wyniku takiego wyboru otrzymuje się H klas a-elementowych /dany obiekt A_r może należeć do kilku klas jednocześnie/.

5. Dla każdej otrzymanej klasy P_h^a oblicza się dyskryminantę według wzoru:

$$\frac{S_h^{P_h} + W_h^{P_h}}{2 + G_h^{P_h}} \quad \text{/dla miar odległości;/} \quad \text{/4.7/}$$

$$\frac{S_h^{P_h} + /1 - W_h^{P_h}/}{2 + /1 - G_h^{P_h}/} \quad \text{/dla miar bliskości;/} \quad \text{/4.8/}$$

gdzie: $W_h^{P_h}$ - miara wewnętrznej spójności klasy P_h /przeciętne podobieństwo wewnątrzklasowe między obiektami danej klasy/;

G^h - miara separowalności klasy P_h /przeciętne podobieństwo obiektów danej klasy do obiektów klas pozostałych/;

S^h - miara stabilności podobieństwa klasy P_h /liczba klas do których należą obiekty danej klasy podzielona przez iloczyn liczby obiektów i liczby klas/.

Wybiera się tę klasę dla której wartość dyskryminanty jest najmniejsza.

6. Rozpatruje się nowy zbiór obiektów, który powstaje przez usunięcie ze zbioru aktualnego obiektów należących do wybranej klasy w kroku 5. Jeśli nowy zbiór nie jest zbiorem pustym to wraca się do kroku 2 /wykreślając z macierzy podobieństw $[C^*]$ wiersze i kolumny odpowiadające indeksom obiektów należących do wybranej w kroku 5 klasy/, w przeciwnym zaś przypadku kończy się postępowanie

Cechą powyższych dwóch metod klasyfikacji jest duża pracochłonność oraz to, że zazwyczaj wyznaczają dużo klas mało licznych.

M E T O D Y D E G L O M E R A C Y J N E

Algorytm postępowania w niehierarchicznych metodach deglomeracyjnych jest następujący:

1. punktem wyjścia jest jedna klasa P obejmująca wszystkie obiekty badania A_1, A_2, \dots, A_n ;
2. w jednym lub kilku krokach dokonuje się podziału klasy P na klasy o mniejszej liczbie obiektów; sposób podziału zależy od stosowanej metody;
3. efektem końcowym podziału jest klasyfikacja $P = \{P_1, P_2, \dots, P_H\}$ o liczbie klas N/P wynikającej z zastosowanej metody.

W literaturze istnieje niewiele metod deglomeracyjnych realizujących powyższy algorytm.

4.3. Wyznaczanie optymalnej liczby klas.

Przedstawione w rozdziale 4.2 metody klasyfikacji nie dają odpowiedzi na pytanie jak wyznaczyć optymalną liczbę klas na które należy podzielić zbiór badanych obiektów A /oprócz metody grafowej $W. Fluty$ /. Problem ten jest niezwykle istotny, jeżeli chodzi o potencjalnego użytkownika metod klasyfikacji, bowiem nie interesuje go zazwyczaj sama zastosowana metoda, lecz podział zbioru obiektów na klasy otrzymany w wyniku jej zastosowania. W literaturze przedmiotu [11; 12; 27; 29; 42; 65; 69; 82] można spotkać wiele propozycji w tym zakresie /świadczą to o ważności tego problemu/, lecz nie ma wśród nich doskonałych. Trwają ciągle poszukiwania sposobów rozwiązania problemu wyznaczania optymalnej liczby klas.

W pracy [65] R. Mojena podaje dwie propozycje wyznaczania optymalnej liczby klas w hierarchicznych metodach aglomeracyjnych. Sposoby te bazują na rozkładzie wartości poziomu połączenia klas $/L/$. Wartości poziomu L rozłożone są monotonicznie rosnąco w przypadku miar odległości oraz monotonicznie malejąco w przypadku miar bliskości. Sposoby te nie znajdują zastosowania dla wyników metod centroidalnych /"mediany" i "środek ciężkości"/ oraz metody średniego połączenia wewnątrz nowej klasy. Wynika to z faktu, że wartości poziomu L związane z łączeniem klas najbardziej podobnych mogą w tych metodach podnosić się i spadać przy przechodzeniu z kroku na krok w klasyfikacji hierarchicznej.

Sposób 1

W sposobie pierwszym na podstawie $n - 1$ wartości poziomu połączenia klas wyznacza się średnią arytmetyczną $/\bar{L}/$ oraz odchylenie standardowe $/S_L/$. Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok $e /e = 1, 2, \dots, n - 2/$ jako pierwszy spełnia nierówność:

$$L_{e+1} > \bar{L} + a S_e \quad \text{/dla miar odległości/,} \quad \text{/4.9/}$$

gdzie: L_{e+1} - wartość poziomu połączenia klas w kroku $e + 1$,
a - dowolna liczba rzeczywista ustalona tak, aby otrzymać nietrywialny podział wynikowy,
e - numer kroku w hierarchicznej procedurze klasyfikacji.

W przypadku, gdy macierz podobieństw [C] była ustalona w oparciu o miarę bliskości to znak nierówności /4.9/ należy zamienić na przeciwny.

Sposób 2

W sposobie drugim również liczy się średnią arytmetyczną oraz odchylenie standardowe z wartości połączenia klas /dodatkowo jeszcze pewne wielkości korygujące/, jednakże wyznacza się je sekwencyjnie po każdym kolejnym e-tym kroku procedury. Z tego tytułu wynika przewaga tego sposobu wyznaczania optymalnej liczby klas nad sposobem pierwszym. Chcąc wyznaczyć optymalną liczbę klas w sposobie pierwszym należy znać wszystkie wartości poziomu połączenia klas /tj. $n - 2$ /, podczas gdy w sposobie drugim tylko $e / e \leq n - 2$ /. Sekwencyjność postępowania w sposobie drugim powoduje, że proces wyznaczania kolejnych podziałów kończy się, gdy spełniona jest relacja /4.10/.

Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok $e / e = e', e'+1, \dots, n - 2$ jako pierwszy spełnia nierówność:

$$L_{e+1} > \bar{L}_e + \beta_e + \gamma_e + a S_e \quad \text{/dla miar odległości/,} \quad \text{/4.10/}$$

gdzie: \bar{L}_e - średnia ruchoma w kroku e /z wartości L_1, \dots, L_e /,
 S_e - odchylenie standardowe ruchome w kroku e /z wartości L_1, \dots, L_e /,

$$\beta_e = \frac{6 \left[2 \sum_{f=e-e'+1}^e w_f L_f - (e'+1) \sum_{f=e-e'+1}^e L_f \right]}{e'/e'^2 - 1},$$

$$\gamma_e = (e' - 1) \beta_e / 2,$$

$$w_f = w_{f-1} + 1; f = e - e' + 2, \dots, e; \text{gdzie: } w_{e-e'+1} = 1.$$

W sytuacji, gdy macierz podobieństw była ustalona w oparciu o miarę bliskości to znak powyższej nierówności należy zamienić na przeciwny.

W hierarchicznych procedurach aglomeracyjnych klasy tworzy się w ten sposób, aby minimalizować stratę informacji towarzyszącą wzrostowi uogólnienia w poszczególnych krokach klasyfikacji. B. J. L. Berry [12] zaproponował procedurę pozwalającą z ciągu klasyfikacji wybrać optymalną. Jako miernik straty informacji Berry zastosował odległość wewnątrzklasową. Celem uzyskania właściwych podziałów konieczne jest przerwanie procesu klasyfikacji na jakimś poziomie straty informacji odpowiadającym określonej iteracji. Problem ten jak dotychczas nie został w pełni rozwiązany. Proponuje się przerwać proces klasyfikacji po tej iteracji, po której następuje wyraźny skok w utracie informacji szczegółowej. Jeśli jest tylko jeden taki skok to wybór jest oczywisty, natomiast przestaje on być oczywisty, gdy nie ma wyraźnego skoku w ogóle lub jest ich kilka.

Innego typu funkcję służącą do ustalenia optymalnej liczby klas zaproponowali Fortier i Solomon [29]. Z ciągu klasyfikacji /otrzymanych hierarchicznymi metodami aglomeracyjnymi lub deglomeracyjnymi/ proponują oni wybrać tę, dla której następująca funkcja:

$$\sum_{P_h^e \in P^e} \left(\sum_{A_r, A_s \in P_h^e} d/A_r, A_s/ - \frac{1}{2} N/P_h^e/ \left[N/P_h^e/ - 1 \right] d^x \right), \quad /4.11/$$

$$\sum_{P_h^e \in P^e} \left(\frac{1}{2} N/P_h^e/ \left[N/P_h^e/ - 1 \right] g^x - \sum_{\substack{A_r, A_s \in P_h^e \\ r < s}} g/A_r, A_s/ \right), \quad /4.12/$$

gdzie: $e = 1, 2, \dots, n - 2$;

P^e - zbiór klas w e -tym kroku procedury hierarchicznej;

P_h^e - h -ta klasa w e -tym kroku procedury hierarchicznej;

$N/P_h^e/$ - liczebność klasy P_h^e w e -tym kroku procedury hierarchicznej;

d^x, g^x - krytyczna wartość miary odpowiednio odległości i bliskości;

osiąga minimum.

W literaturze przedmiotu spotkać można jeszcze wiele metod wyszukiwania klasyfikacji optymalnej z ciągu klasyfikacji, jednakże nie wnoszą one nic nowego w stosunku do wyżej omówionych lub różnią się od nich nieznacznie /por. [43; 69; 82] /.

Metoda taksonomii wrocławskiej /a co za tym idzie i metoda Prima/ zawiera pewne sposoby pozwalające z ciągu klasyfikacji wybrać optymalną. Mając ustawione w dendrycie podobieństwa między sąsiadującymi obiektami /graficznie reprezentowane przez krawędzie/ w porządku wzrastającego podobieństwa /malejących wartości odległości d_e lub wzrastających wartości bliskości g_e / oblicza się ilorazy sąsiednich wyrazów [27; 72] :

$$w_e = \frac{d_e}{d_{e+1}} , \quad /e, e' = 1, 2, \dots, n - 2/, \quad /4.13/$$

$$w_e = \frac{g_{e+1}}{g_e} \quad . \quad /4.14/$$

Z otrzymanych podziałów na $N/P^e/$ i $N/P^{e+1}/$ klas wybiera się ten dla którego zachodzi /otrzymuje się wtedy podział naturalny/:

$$w_e < w_{e+1} \quad . \quad /4.15/$$

Z dwóch podziałów naturalnych na $N/P^e/$ i $N/P^{e+1}/$ klas ten jest lepszy, który ma mniejszą wartość w_e .

Inny sposób wyznaczania optymalnej liczby klas prezentuje Z. Hellwig w pracy [40]. Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok e / $e = 1, 2, \dots, n - 2/$ jako ostatni spełnia nierówność:

$$d_e > \bar{d} + 2 S_d \quad , \quad /4.16/$$

$$g_e < \bar{g} + 2 S_g \quad , \quad /4.17/$$

gdzie: d_e, g_e - wartość miary odległości /odpowiednio bliskości/ między obiektami sąsiadującymi w dendrycie w kroku e -tym,

$$\bar{d} = \frac{1}{n} \sum_s \min_r d/A_r, A_s \quad .$$

$$S_d = \left[\frac{1}{n} \sum_s / \min_r d/A_r, A_s / - \bar{d} /^2 \right]^{0,5} \quad ,$$

analogiczne wzory otrzymuje się dla miary bliskości wstawiając w miejsce symbolu d symbol g .

Słabością większości metod niehierarchicznych jest problem doboru różnych parametrów, które decydują w znacznej mierze o ostatecznym wyniku klasyfikacji. Parametry te dobierane są zazwyczaj arbitralnie przez badacza. Można wskazać na pewne propozycje /są one wykorzystywane w znacznej liczbie metod klasyfikacji/ w sprawie ustalania parametru ρ podane przez autorów pracy [11]:

$$\varrho = \max_s \min_r \{d/A_r, A_s/\} , \quad /4.18/$$

$$\varrho = \min_s \max_r \{g/A_r, A_s/\} , \quad /4.19/$$

oraz

$$\varrho = \bar{d} + a S_d , \quad /4.20/$$

$$\varrho = \bar{g} + a S_g , \quad /4.21/$$

gdzie: \bar{d} , \bar{g} , S_d oraz S_g wyznacza się tak jak we wzorach /4.16/ i /4.17/.

5. REJESTRACJA ZMIAN W CZASIE W WYNIKACH KLASYFIKACJI

5.1. Uwagi wstępne.

W rozdziale tym omówione zostaną pewne miary służące do porównania wyników klasyfikacji zbioru obiektów otrzymanych na podstawie [100] :

- a/ dwóch różnych metod klasyfikacji,
- b/ dwóch różnych zestawów cech /za pomocą jednej metody klasyfikacji/,
- c/ informacji statystycznych pochodzących z dwóch różnych okresów /za pomocą tej samej metody klasyfikacji i tego samego zespołu cech/.

Dalsze rozważania ograniczone zostaną do sposobów rejestracji zmian w wynikach klasyfikacji obiektów uzyskanych na podstawie informacji pochodzących z dwóch różnych okresów /patrz punkt c/, choć znajdują one zastosowanie również przy porównywaniu wyników klasyfikacji w rozumieniu punktu a i b. Omawiane w tym rozdziale sposoby rejestracji zmian w czasie w otrzymanych wynikach klasyfikacji mają charakter uniwersalny, tzn. mogą być zastosowane nie tylko w badaniach strukturalnych.

Zastosowanie jednej z metod klasyfikacji omówionych w rozdziale czwartym - do podziału zbioru obiektów - daje wyniki w postaci ciągu klasyfikacji /metody hierarchiczne/ lub jednego podziału wynikowego /metody niehierarchiczne/.

5.2. Mierzenie podobieństwa między ciągami klasyfikacji.

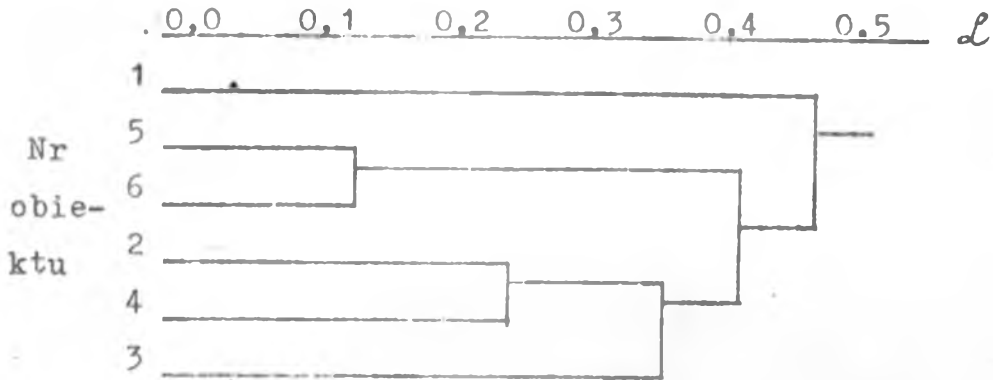
W tym miejscu omówione zostaną sposoby mierzenia podobieństwa między ciągami klasyfikacji otrzymanymi przez zastosowanie hierarchicznych metod aglomeracyjnych. Graficzną prezentacją hierarchicznych metod aglomeracyjnych są tzw. dendrogramy. W oparciu o nie konstruuje się przekształcone macierze podobieństw $[C_0^x]$ i $[C_1^x]$, których porównanie pozwala ocenić zmiany jakie zaszły w badanych okresach w hierarchicznych ciągach klasyfikacji. Każdy element przekształconej macierzy podobieństw $[C^x]$ informuje na jakim poziomie połączenia klas \mathcal{L} dane obiekty zostały włączone do jednej klasy. Przekształcone macierze podobieństw mogą być porównywane za pomocą następującej miary:

$$f/C_1^x; C_0^x/ = \frac{\sum_{r=2}^{r=n} \sum_{s=1}^{s=r-1} |c_1^x/A_r, A_s/ - c_0^x/A_r, A_s/|}{\binom{n}{2}} \quad /5.1/$$

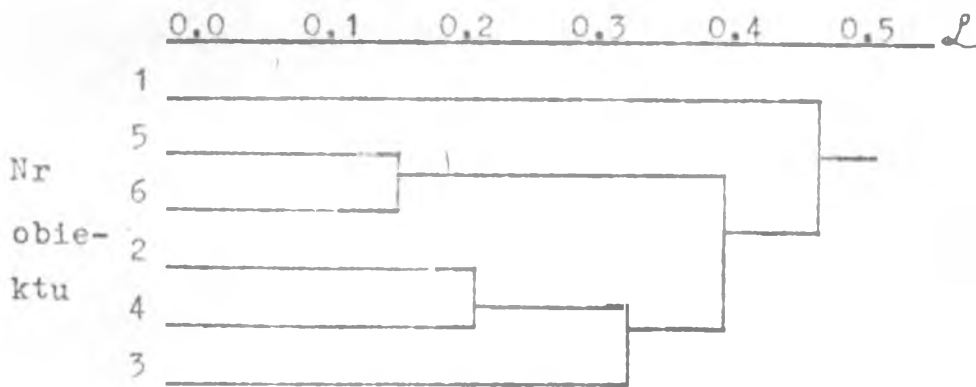
Z racji tego, że miary podobieństwa struktur, a co za tym idzie i podobieństwa międzyklasowe /por. rozdział 3.1 i 4.2.1/ zawarte są w przedziale $\langle 0; 1 \rangle$ to miara o zapisie /5.1/ również przyjmuje wartości z tego przedziału.

W sytuacji, gdy miara przyjmuje wartość 0 oznacza to, że nie zaszły żadne zmiany w porównywanych ciągach klasyfikacji. Z drugiej strony, gdy przyjmuje wartość 1, to wszystkie unormowane obrazy struktur obiektów są identyczne w okresie T_0 , natomiast zupełnie różne w okresie T_1 . Oczywiście wszystkie unormowane obrazy struktur obiektów mogą być zupełnie różne wtedy, gdy $n \leq k$. W pozostałych przypadkach $\langle n \rangle > k$ miara powyższa będzie przyjmować wartości z przedziału lewostronnie domkniętego $\langle 0; 1 \rangle$.

Rozważania te można zilustrować prostym przykładem. Dendrogramy 6 obiektów w dwóch okresach T_0 i T_1 oraz odpowiadające im macierze $[C^x]$ przedstawia rys. 5.1.



a/ dendrogram w okresie T_0



b/ dendrogram w okresie T_1

2	0,48				
3	0,48	0,36			
4	0,48	0,24	0,36		
5	0,48	0,41	0,41	0,41	
6	0,48	0,41	0,41	0,41	0,13
	1	2	3	4	5

2	0,48				
3	0,48	0,32			
4	0,48	0,21	0,32		
5	0,48	0,39	0,39	0,39	
6	0,48	0,39	0,39	0,39	0,16
	1	2	3	4	5

a/ przekształcona macierz podobieństw $\begin{bmatrix} C^x \\ 0 \end{bmatrix}$ b/ przekształcona macierz podobieństw $\begin{bmatrix} C^x \\ 1 \end{bmatrix}$

Rys. 5.1. Dendrogramy 6 obiektów otrzymane metodą pojedynczego połączenia oraz odpowiadające im przekształcone macierze podobieństw w okresach T_0 i T_1 .

Po zastosowaniu wzoru /5.1/ otrzymuje się wartość 0,017 oznaczającą, że w porównywanych okresach nastąpiły niewielkie zmiany w ciągach klasyfikacji.

Innego typu propozycje miar można znaleźć w pracach [4; 17; 37; 90]. W jednej z nich do pomiaru podobieństwa między macierzami $\begin{bmatrix} C^x \\ C_0^x \end{bmatrix}$ i $\begin{bmatrix} C^x \\ C_1^x \end{bmatrix}$ proponuje się przyjąć współczynnik korelacji kofenetycznej /jest to współczynnik korelacji liniowej według momentu iloczynowego między odpowiadającymi sobie elementami w przekształconych macierzach podobieństw w okresach T_0 i T_1 /. Posiada on jednak tę wadę, że jest niezmienny względem liniowej transformacji macierzy $\begin{bmatrix} C^x \\ C_0^x \end{bmatrix}$. Oznacza to, że korelacja nie ulega zmianie wtedy, gdy dodamy do każdego elementu macierzy $\begin{bmatrix} C^x \\ C_0^x \end{bmatrix}$ stałą lub każdy element pomnożymy przez stałą. Przekształcone macierze podobieństw $\begin{bmatrix} C^x \\ C_0^x \end{bmatrix}$ i $\begin{bmatrix} C^x \\ C_1^x \end{bmatrix}$ będą do siebie podobne wtedy, gdy różnice między odpowiadającymi sobie elementami w tych macierzach będą niewielkie. Duża wartość współczynnika korelacji nie oznacza wcale, że porównywane ciągi klasyfikacji są do siebie bardzo podobne, ponieważ przekształcona macierz podobieństw $\begin{bmatrix} C^x \\ C_0^x \end{bmatrix}$ ze stosunkowo małymi liczbami może być silnie skorelowana z macierzą $\begin{bmatrix} C^x \\ C_1^x \end{bmatrix}$ zawierającą stosunkowo duże liczby.

5.3. Mierzenie podobieństwa między podziałami wynikowymi.

Odmiennych sposobów pomiaru podobieństwa należy szukać wtedy, gdy efekt końcowy klasyfikacji przedstawiony jest w postaci jednego podziału wynikowego, a nie w formie ciągu klasyfikacji. Jeden podział otrzymuje się w efekcie zastosowania metod niehierarchicznych oraz w rezultacie wyboru z ciągu klasyfikacji podziału optymalnego według zadanego z góry kryterium.

W takim układzie jednym ze sposobów badania podobieństwa między dwoma podziałami jest konstrukcja tablicy kontyngencji /tab. 5.1/, celem stworzenia klasyfikacji skrzyżowanej dwóch podziałów

[4; 31; 33; 85] . Na podstawie tablicy kontyngencji można uzyskać następujące informacje:

1. Jaka jest liczba klas w obydwu podziałach /dane te zawarte są w główce i boczku tablicy kontyngencji/.
2. Jaka jest liczebność obiektów w klasach w obydwu podziałach /dane te uzyskuje się analizując rozkłady brzegowe/.
3. Pod warunkiem znajomości składu klasy P_h w jednym podziale można dowiedzieć się ile obiektów spośród tej klasy należy do klas P_1, P_2, \dots, P_H w drugim podziale i odwrotnie. Informacje te uzyskuje się analizując rozkłady warunkowe.
4. Wnętrze tablicy mówiące o rozkładzie łącznym informuje o liczbie obiektów, które jednocześnie należą do jednej z klas podziału z okresu T_0 i T_1 .
5. Jaki jest stopień podobieństwa między porównywanymi podziałami.

Tab. 5.1

Tablica kontyngencji /klasyfikacja skrzyżowana dwóch podziałów/

Podział w T_1		klasy				$\sum_h n_{h'h}$
		P_1	P_2	...	P_H	
Podział w T_0	P_1	n_{11}	n_{12}	...	n_{1H}	$n_{1.}$
	P_2	n_{21}	n_{22}	...	n_{2H}	$n_{2.}$
klasy	•	•	•	•	•	•
	•	•	•	•	•	•
	•	•	•	•	•	•
	P_H	$n_{H'1}$	$n_{H'2}$...	$n_{H'H}$	$n_{H'.$
$\sum_P n_{h'h}$		$n_{.1}$	$n_{.2}$...	$n_{.H}$	$n_{..}$

gdzie: $h' = 1, 2, \dots, H'$; $h = 1, 2, \dots, H$.

Informacje płynące z pierwszych czterech punktów są wykorzystywane w punkcie piątym przy konstrukcji syntetycznych miar rejestrujących zmiany w wynikach klasyfikacji w porównywanych okresach.

W literaturze przedmiotu [4; 33; 99; 100] dużo miejsca poświęca się przedstawianiu różnych miar podobieństwa dwóch podziałów, natomiast nie zwraca się uwagi na własności, jakie powinna spełniać dobrze określona miara. Zdaniem autora dobrze zdefiniowana miara podobieństwa, rejestrująca zmiany w czasie w wynikach klasyfikacji, powinna spełniać następujące wymogi [113] :

1^o Winna być łatwa w interpretacji, co prowadzi do postulatu, aby jej wartości były zawarte w zamkniętym przedziale liczbowym. Wygodne jest, z praktycznego punktu widzenia, aby miara przyjmowała wartości z przedziału $\langle 0; 1 \rangle$:

$$0 \leq f/P^{1/}; P^{0/} \leq 1,$$

gdzie: $P^{1/} = \{P_1^{1/}, P_2^{1/}, \dots, P_H^{1/}\}$ - klasyfikacja w okresie T_1 ,

$P^{0/} = \{P_1^{0/}, P_2^{0/}, \dots, P_H^{0/}\}$ - klasyfikacja w okresie T_0 ,

f - postać analityczna funkcji podobieństwa.

2^o

$$f/P^{1/}, P^{0/} = \begin{cases} 0 & \text{dla miar odległości} \\ 1 & \text{dla miar bliskości} \end{cases}$$

wtedy i tylko wtedy, gdy porównywane klasyfikacje są identyczne, tzn. $P^{1/} = P^{0/}$. Sytuacja taka ma miejsce w przypadku kwadratowej tablicy kontyngencji, w której każdej klasie pierwszego podziału odpowiada tylko jedna klasa w podziale drugim.

3^o

$$f/P^{1/}, P^{0/} = \begin{cases} 1 & \text{dla miar odległości} \\ 0 & \text{dla miar bliskości} \end{cases}$$

wtedy i tylko wtedy, gdy jeden podział zawiera tyle klas ile jest

obiektów, a drugi jedną klasę zawierającą wszystkie obiekty.

4° Warunek symetrii:

$$f/P^{1/}, P^{0/}/ = f/P^{0/}, P^{1/}/ .$$

5° Wartość miary jest niezmienna względem jakiegokolwiek uporządkowania wierszy lub kolumn w tablicy kontyngencji.

6° Koncepcja oraz postać analityczna miary powinny być wystarczająco proste, żeby nie sprawiały trudności numerycznych.

Do pomiaru podobieństwa między dwoma podziałami pochodzącymi z dwóch okresów proponuje się wykorzystać miary zależności statystycznej [4; 33] .

Najstarsze i tradycyjne miary współzależności oparte są na statystyce χ^2 określonej wzorem:

$$\chi^2 = \sum_h \sum_{h'} /n_{h'h} - \hat{n}_{h'h} /^2 / \hat{n}_{h'h} , \quad /5.2/$$

gdzie: $n_{h'h}$ i $\hat{n}_{h'h}$ to empiryczne i teoretyczne /hipotetyczne/ liczebności tablicy kontyngencji.

Zakładając prawdziwość hipotezy zerowej /mówiącej o niezależności obydwóch podziałów/ liczebności hipotetyczne wyznacza się ze wzoru:

$$\hat{n}_{h'h} = \frac{n_{.h} n_{h'.}}{n_{..}} . \quad /5.3/$$

Po przejściu z rozkładu liczebności na rozkład częstości stosując wzory:

$$f_{h'.} = \frac{n_{h'.}}{n_{..}} ; f_{.h} = \frac{n_{.h}}{n_{..}} ; f_{h'h} = \frac{n_{h'h}}{n_{..}} ; \quad /5.4/$$

otrzymuje się

$$\chi^2 = n_{..} \sum_h \sum_{h'} /f_{h'h} - f_{h'.} f_{.h} /^2 / f_{h'.} f_{.h} . \quad /5.5/$$

Statystyka ta ma, przy założeniu prawdziwości hipotezy zerowej, rozkład χ^2 z $/H - 1/H' - 1/$ stopniami swobody. Wartość krytyczną χ_c^2 odczytuje się z tablic rozkładu χ^2 dla ustalonego z góry poziomu istotności \mathcal{L} i $/H - 1/H' - 1/$ stopni swobody. Hipotezę zerową należy odrzucić, gdy zachodzi nierówność $\chi^2 \gg \chi_c^2$ co oznacza, że występuje zależność obu podziałów. W przeciwnym przypadku nie ma podstaw do odrzucenia hipotezy o niezależności podziałów wynikowych.¹

W związku z tym, że statystyka χ^2 informuje tylko o fakcie występowania współzależności, a nie informuje o sile tej zależności wprowadzono miary oparte na statystyce χ^2 .

Do tej grupy miar można zaliczyć takie jak współczynnik średniej kwadratowej wielodzzielczości ϕ^2 , współczynnik Czuprowa T^2 , współczynnik Cramera C^2 i współczynnik Pearsona P^2 określone odpowiednio wzorami [18; 22; 31]:

$$\phi^2 = \chi^2 / n_{..} \quad ; \quad T^2 = \frac{\chi^2 / n_{..}}{\{ /H - 1/H' - 1/ \}^{0,5}} \quad ; \quad /5.6 - 5.7/$$

$$C^2 = \frac{\chi^2 / n_{..}}{\min /H, H' / - 1} \quad ; \quad P^2 = \frac{\chi^2}{n_{..} + \chi^2} \quad . \quad /5.8 - 5.9/$$

Spośród innych miar współzależności warto odnotować propozycję Z. Hellwiga [38, s.143]:

$$H^2 = \frac{1 - \sum_h \sum_{h'} \min /f_{h'h}; f_{h'}. f_{.h}/}{1 - 1 / \min /H, H' /} \quad . \quad /5.10/$$

¹Rozważania te mają sens tylko wtedy, gdy zbiór badanych obiektów A stanowi próbę pobraną z pewnej populacji generalnej \mathcal{A} .

W tym miejscu wykazana zostanie znikoma przydatność miar zależności do rozwiązania problemu określonego w tytule rozdziału. Miary zależności nie spełniają wszystkich postulatów określonych w tym rozdziale dla miar podobieństwa między dwoma podziałami z okresu T_0 i T_1 /por. tab. 5.2/.

Tab. 5.2

Własności miar zależności

Postulaty nr wzoru	1 ^o	2 ^o	3 ^o	4 ^o	5 ^o
/5.6/	-	-	-	+	+
/5.7/	+	+	-	+	+
/5.8/	+	-	-	+	+
/5.9/	-	-	-	+	+
/5.10/	+	+	-	+	+

+ spełnia postulat; - nie spełnia postulatu.

Źródło: opracowanie własne.

Wszystkie formuły spełniają z jednej strony postulat symetryczności i niezmienności względem jakiegokolwiek uporządkowania wierszy lub kolumn, natomiast z drugiej strony nie spełniają warunku 3^o. Wynika to z faktu, że niezależność statystyczna dwóch podziałów nie pociąga za sobą zupełnego niepodobieństwa tychże podziałów.

Miara zależności Cramera nie spełnia warunku 2^o, ponieważ możliwe są inne sytuacje, w których miara ta przyjmuje wartość 1. Górna granica tej miary jest osiągnięta wtedy i tylko wtedy, gdy każdy wiersz $/H \succcurlyeq H/$ lub każda kolumna $/H \preccurlyeq H/$ zawiera tylko jeden element różny od zera. Jak z tego widać miara ta może przyjmować wartość 1 nie tylko dla tablic kwadratowych. Warunku 2^o nie speł-

nią również miary ϕ^2 i P^2 . Pierwsza z nich może przyjmować wartości większe od jedności, a druga z kolei nigdy nie osiąga górnej granicy, dążąc do niej w miarę zwiększania liczby wierszy i kolumn w tablicy kontyngencji [33].

Ponadto wszystkie miary zależności statystycznej są nieokreślone, gdy w jednym z podziałów jest tylko jedna klasa zawierająca wszystkie obiekty.

Miarę podobieństwa dwóch podziałów, która opiera się również na tablicy kontyngencji, zaproponował C. Szmigiel w pracy [100]:

$$S^{/1/} = \min /S_1, S_2/ , \quad /5.11/$$

gdzie:

$$S_1 = \frac{\sum_h \max_h n_{h'h} - \max /H'; \max_h n_{.h}/}{n_{..} - \max /H'; \max_h n_{.h}/} , \quad /5.12/$$

$$S_2 = \frac{\sum_h \max_h n_{h'h} - \max /H; \max_{h'} n_{h'.}/}{n_{..} - \max /H; \max_{h'} n_{h'.}/} . \quad /5.13/$$

W pracy [99] tenże sam autor przedstawia modyfikacje proponowanej miary podobieństwa dwóch podziałów:

$$S^{/2/} = \max /S_1; S_2/ , \quad /5.14/$$

$$S^{/3/} = \frac{1}{2} /S_1 + S_2/ , \quad /5.15/$$

$$S^{/4/} = /S_1 S_2/^{0,5} . \quad /5.16/$$

Na podstawie badań symulacyjnych C. Szmigiel dochodzi do wniosku, że najlepszą przydatność praktyczną ma miara $S^{/2/}$, przede wszystkim ze względu na dużą zmienność.

Konstrukcja miar globalnych utworzonych z miar cząstkowych

S_1 i S_2 oraz występowanie w licznikach miar cząstkowych liczby wydzielonych klas w obu podziałach $/H'$; $H/$ powoduje, że miary C. Szmigla są nieokreślone w następujących dwóch sytuacjach [113] :

- a/ gdy w jednym z podziałów jest tylko jedna klasa zawierająca wszystkie obiekty,
- b/ gdy liczba klas w jednym z podziałów jest równa liczebności zbioru obiektów.

W pozostałych przypadkach miary te przyjmują wartości z przedziału obustronnie domkniętego $\langle 0; 1 \rangle$. Wartość 0 przyjmują w przypadku statystycznej niezależności obu podziałów oraz wtedy, gdy istnieje a , b takie, że

$$n_{h'a} = \max_h n_{h'h} \quad /dla\ wszystkich\ h' /$$

lub /i - dla miar $S^{/2/}$ i $S^{/3/}$ /

$$n_{bh} = \max_{h'} n_{h'h} \quad /dla\ wszystkich\ h /$$

Z kolei wartość 1 przyjmują, jeżeli każdy wiersz i /lub - dla miary $S^{/2/}$ / kolumna zawiera tylko jeden niezerowy element $n_{h'h}$.

Wszystkie miary zaproponowane przez C. Szmigla nie spełniają warunku 3^o, natomiast warunek 2^o spełniają miary $S^{/1/}$, $S^{/3/}$ i $S^{/4/}$ przy założeniu, że są określone.

Miary cząstkowe /5.12/ i /5.13/ zaproponowane przez C. Szmigla są analogiczne do miar podanych w pracy [31] przez L. A. Goodmana i W. H. Kruskala /por. [113] /:

$$\lambda_A = \frac{\sum_{h'} \max_h n_{h'h} - \max_h n_{.h}}{n_{..} - \max_h n_{.h}}, \quad /5.17/$$

$$\lambda_B = \frac{\sum_h \max_{h'} n_{h'h} - \max_{h'} n_{h'.}}{n_{..} - \max_{h'} n_{h'.}} \quad /5.18/$$

Każda z miar cząstkowych z osobna nie spełnia warunku symetryczności, przeto autorzy podają ogólną formę miary spełniającą ten warunek:

$$\Lambda = \frac{\sum_h \max_h n_{h'h} + \sum_{h'} \max_{h'} n_{h'h} - \max_h n_{\cdot h} - \max_{h'} n_{h' \cdot}}{2n_{\cdot \cdot} - \max_h n_{\cdot h} - \max_{h'} n_{h' \cdot}} \quad /5.19/$$

Miary cząstkowe Goodmana-Kruskala różnią się od miar C. Szmigla tym, że w ich liczniku nie występuje element $/H'$ lub $H/$, który oznacza liczbę wyodrębnionych klas. Powoduje to, że miara Λ z tego tytułu nie posiada jednej z wad miar C. Szmigla, a mianowicie Λ jest określona wtedy, gdy liczba klas w jednym z podziałów jest równa liczebności zbioru obiektów /patrz punkt b przy opisie miar C. Szmigla/. Ponadto z konstrukcji miary Λ wynika, że jest ona nieokreślona tylko wtedy, gdy tablica kontyngencji ma wymiary 1×1 , tzn. w obu klasyfikacjach występuje jedna klasa obejmująca wszystkie obiekty badania. W pozostałych przypadkach miara Λ zawiera się w przedziale $\langle 0; 1 \rangle$. Wartość 1 przyjmuje wtedy, gdy każdy wiersz i każda kolumna zawiera tylko jeden niezerowy element $n_{h'h}$. Z kolei wartość 0 przyjmuje w przypadku statystycznej niezależności obu podziałów oraz wtedy, gdy istnieje a, b takie, że

$$n_{h'a} = \max_h n_{h'h} \quad /dla\ wszystkich\ h'/$$

i

$$n_{bh} = \max_{h'} n_{h'h} \quad /dla\ wszystkich\ h/$$

Miara Λ podobnie jak miary C. Szmigla nie spełnia warunku 3^o, natomiast warunek 2^o spełnia przy założeniu, że jest określona.

Poniżej zaprezentowane zostaną inne miary podobieństwa między dwoma podziałami spełniające warunki 1^o - 6^o określone dla tychże miar.

Zasadnicze znaczenie ma w tej grupie miara podobieństwa dwóch podziałów zaproponowana przez W. M. Randa [83], a określona wzorem:

$$R = \sum_{r=2}^{r=n} \sum_{s=1}^{s=r-1} \gamma_{rs} / \binom{n}{2}, \quad /5.20/$$

gdzie:

$$\gamma_{rs} = \begin{cases} 1 & , \text{ jeżeli obiekty } A_r \text{ i } A_s \text{ w obu podziałach znajdują się w tych samych klasach,} \\ 1 & , \text{ jeżeli obiekty } A_r \text{ i } A_s \text{ w obu podziałach znajdują się w różnych klasach,} \\ 0 & , \text{ w pozostałych przypadkach.} \end{cases}$$

W przypadku dużej liczby obiektów wykorzystanie tej miary byłoby bardzo pracochłonne /w sposobie tym rozpatruje się $\frac{1}{2} n/n - 1/$ kombinacji par obiektów/, zatem Rand proponuje efektywniejszą /w sensie czasu pracy/ postać tej miary, która opiera się na tablicy kontyngencji:

$$R = \left[\binom{n}{2} - \left[\frac{1}{2} \left(\sum_h n_{\cdot h}^2 + \sum_{h'} n_{h' \cdot}^2 \right) - \sum_h \sum_{h'} n_{h' h}^2 \right] \right] / \binom{n}{2}. \quad /5.21/$$

Przedział zmienności tej miary zaczyna się od 0, kiedy to dwa podziały są zupełnie niepodobne /jeden podział zawiera tyle klas ile jest obiektów, a drugi jedną klasę zawierającą wszystkie obiekty/, a kończy na 1, kiedy podziały są identyczne.

Identyczne kryterium podobieństwa podziałów zastosował w konstrukcji miary A. Sokołowski [69; 81], otrzymując w efekcie postać analogiczną do miary Randa:

$$R_S = 1 - \sum_{r=2}^{r=n} \sum_{s=1}^{s=r-1} (1 - \gamma_{rs}) / \binom{n}{2} = R. \quad /5.22/$$

Podobieństwo między dwoma podziałami można ustalić przy pomocy miar skojarzenia opartych na tablicy o wymiarach 2 x 2 /por. tab. 5.3/.

Tab. 5.3

Tablica 2 x 2

Podział w T_1	1	0	Ogółem
Podział w T_0	1	0	Ogółem
1	a	b	a + b
0	c	d	c + d
Ogółem	a + c	b + d	a + b + c + d

gdzie: a, b, c, d - liczebności obiektów w klasach.

W sposobie tym rozpatruje się $\frac{1}{2} n/n - 1/$ kombinacji par obiektów, które następnie zakwalifikowane zostają w każdym podziale do dwóch klas:

klasa 0 : obiekty A_r i A_s znajdują się w różnych klasach w podziale,

klasa 1 : obiekty A_r i A_s znajdują się w tej samej klasie w podziale.

Do porównań między podziałami można użyć dwie spośród binarnych miar skojarzenia przedstawionych przez M. R. Anderberga w tab. 4.5 /por. [4, s.89] /. Najbardziej oczywistą miarą podobieństwa jest tzw. prosta miara skojarzenia o postaci:

$$\frac{a + d}{a + b + c + d} \quad . \quad /5.23/$$

Miara Randa dla tablicy 2 x 2 przyjmuje właśnie powyższą postać, przeto interpretacja jej jest już znana.

Drugą miarą podobieństwa jest współczynnik Rogersa-Tanimoto:

$$\frac{a + d}{a + d + 2/b + c/} \quad . \quad /5.24/$$

w którym dwa razy wyższą wagę otrzymują pary niezgodne w obu podziałach. Pozostałe binarne miary skojarzenia nie spełniają wszystkich postulatów, które wcześniej określono dla miar podobieństwa dwóch podziałów.

Ze względu na to, że miary skojarzenia wymagają rozpatrzenia $\frac{1}{2} n/n - 1/$ kombinacji par obiektów - co jest bardzo pracochłonne autor proponuje, aby do badania podobieństwa podziałów stosować miarę W. M. Randa o postaci /5.21/.

6. ZASTOSOWANIE METOD KLASYFIKACJI DO BADANIA STRUKTURY GAŁĘZIOWEJ PRZEMYSŁU WOJEWÓDZTW POLSKI W ROKU 1975 I 1980

6.1. Formalizacja opisu struktur gałęziowych.

Empiryczna część pracy poświęcona zostanie wykorzystaniu metod klasyfikacji do analizy przestrzennego zróżnicowania struktur gałęziowych przemysłu w Polsce.

Badanie strukturalne rozpoczyna się od konstrukcji macierzy danych /patrz rozdział drugi/, która to wymaga określenia trzech zbiorów, a mianowicie zbioru obiektów badania, zbioru cech strukturalnych oraz zbioru elementów struktury.

Obiektami badania przyjętymi w pracy są województwa Polski wyodrębnione jako jednostki administracyjne 1.06.1975 r. /tworzą one zbiór 49. elementowy/.

Określenie struktura gałęziowa przemysłu wskazuje, że elementami struktury są gałęzie przemysłu. Jest to jednak sformułowanie mało precyzyjne. Otóż elementami struktury będą nie gałęzie przemysłu lecz następujące grupy gałęzi przemysłu /zgodnie z klasyfikacją gospodarki narodowej obowiązującą od 1.01. 1976 r./:

1. Przemysł paliwowo-energetyczny.
2. Przemysł metalurgiczny.
3. Przemysł elektromaszynowy.
4. Przemysł chemiczny.
5. Przemysł mineralny.
6. Przemysł drzewno-papierniczy.
7. Przemysł lekki.

8. Przemysł spożywczy.

9. Pozostałe gałęzie przemysłu.

Przyjęcie grup gałęzi przemysłu zamiast gałęzi przemysłu wynika z tego, że klasyfikacja przemysłu ulega ciągłym zmianom w czasie, co sprawia że traci się porównywalność w układzie gałęziowym. Porównywalność taką zapewniają grupy gałęzi przemysłu, ponieważ zmiany w klasyfikacji przemysłu dokonują się na ogół wewnątrz grup gałęzi przemysłu.

W dalszej części tego rozdziału poruszone zostaną dwa zagadnienia z zakresu struktur gałęziowych przemysłu:

a/ klasyfikacja województw Polski ze względu na poziom uprzemysłowienia,

b/ klasyfikacja województw Polski ze względu na profil /typ/ uprzemysłowienia.

Poziom oraz profil uprzemysłowienia to dwa zagadnienia tego samego pojęcia, którym jest uprzemysłowienie. Przez uprzemysłowienie - ogólnie rzecz biorąc - rozumie się "określoną koncentrację albo nasycenie przestrzeni przemysłem" [101] .

Realizacja powyższych dwóch zagadnień z zakresu struktur gałęziowych wymaga doboru odpowiedniego zestawu cech w klasyfikacji wielocechowej lub ustalenia priorytetu cechy w klasyfikacji jednocechowej. W tym miejscu należy wyjaśnić dlaczego oprócz badania wielocechowego przeprowadzono badanie jednocechowe. Cóż w literaturze /por. np. [14; 71; 97; 101; 107] / jak dotychczas przy badaniu struktur gałęziowych przemysłu posługiwano się tylko badaniem jednocechowym. Praca niniejsza stanowi więc próbę badań wielocechowych w tym zakresie. Badanie jednocechowe przeprowadzono celem jego porównania z wynikami badania wielocechowego. Przeprowadzona poniżej analiza merytoryczna cech wskazuje na przewagę badań wielocechowych nad badaniami jednocechowymi.

Spośród zaprezentowanych w rozdziale drugim podejść do zaga-

dnienia doboru cech zostanie wykorzystane te, które opiera się wyłącznie na doborze merytorycznym. Uzasadnienie takiego postępowania można streścić w kilku punktach, a mianowicie:

- a/ w literaturze można spotkać wiele prac [67; 71; 86; 87; 97; 101; 102; 121] traktujących o tym zagadnieniu, w których występuje względna zgodność co do proponowanych mierników uprzemysłowienia;
- b/ lista potencjalnych cech nie jest zbyt długa;
- c/ lista potencjalnych cech zostaje ponadto ograniczona przez niedostępność danych statystycznych.

W literaturze [14; 55; 71; 97; 122] jako miernik uprzemysłowienia w przypadku badania jednocechowego proponuje się przyjąć jeden z następujących:

- liczba zatrudnionych w przemyśle,
- wartość produkcji globalnej w przemyśle,
- wartość produkcji czystej w przemyśle,
- wartość środków trwałych w przemyśle.

Wszystkie powyższe mierniki mogą być wyrażone w formie relatywnej przez odniesienie ich do liczby ludności lub powierzchni terenu.

Ustalenie priorytetu cechy spośród wyżej wymienionych będzie polegało na merytorycznej analizie określonych zalet i wad tych mierników. Własności poszczególnych mierników zostaną omówione na podstawie prac [55; 86; 97; 122] .

Spośród mierników uprzemysłowienia w literaturze priorytet przyznaje się liczbie zatrudnionych. Miernik ten jest jednoznaczny, porównywalny i wymierny. Wynika to m. in. z łatwości jego ewidencjonowania, które dokonywane jest przecież w jednostkach naturalnych. Fakt ten powoduje, że ów miernik w mniejszym stopniu niż inne obarczony jest błędami pomiaru. Nie jest on jednak wolny od wad. Wadą zatrudnienia jest m. in. pomijanie poziomu kwalifikacji i wydajności siły roboczej, zacieranie różnic między pracownikami bezpośrednio produkcyjnymi i pośrednio produkcyjnymi . Z kolei

cenną zaletą tego miernika jest niezależność od zmian cen.

Własności tej nie posiadają żadne z innych używanych powszechnie mierników /wartość produkcji globalnej, wartość produkcji czystej, wartość produkcji netto, wartość środków trwałych/. Liczenie produkcji globalnej metodą zakładową powoduje, że zawyża się faktyczną wartość produkcji przez kilkakrotne zaliczanie do niej kosztów materiałów, surowców oraz usług obcych. Mniejsze deformacje występują przy posługiwaniu się miernikiem produkcji globalnej liczonej metodą gałęziową, przy której eliminuje się obroty w ramach poszczególnych gałęzi. Z punktu widzenia badań strukturalnych zaletą omawianego miernika jest kompleksowe ujęcie nakładów i efektów procesu produkcyjnego w poszczególnych gałęziach. Praktyczne wykorzystanie tego miernika jak i miernika produkcji czystej oraz netto jest niemożliwe ze względu na brak danych statystycznych /nie prowadzi się dla nich szczegółowej statystyki/. Poprawniejszym miernikiem od produkcji czystej jest produkcja netto, która ujmuje wartość przeniesioną przez zużyte środki produkcji /amortyzację/. Nieujmowanie amortyzacji powoduje zaniżanie produkcji tej gałęzi, w której postęp techniczny przejawia się najsilniej.

Kolejnym miernikiem stosowanym do badania uprzemysłowienia jest wartość środków trwałych. Między teorią ekonomii a praktyką gospodarczą występuje rozbieżność w definiowaniu tego pojęcia. Teoria ekonomii przyjmuje za kryterium różnicowania środków produkcji na trwałe i nietrwałe liczbę cykli produkcyjnych, w których one uczestniczą [122]. Praktyka gospodarcza natomiast za kryterium przyjmuje okres eksploatacji oraz wartość środków produkcji /obie te wielkości reguluje norma prawna/. Co pewien okres następuje regulacja norm prawnych w tym zakresie i zmienia się dolna granica wartości środków produkcji, przy której to jeszcze uznaje się dany środek za trwały. Wymaga rozstrzygnięcia kwestia czy za podstawę szacunku przyjmować wartość początkową /pierwotną/

czy też odtworzeniową środków trwałych. Wartość początkowa brutto środków trwałych jest równa sumie nakładów poniesionych na ich zakup lub wytworzenie, bez potrącenia wartości zużycia. Z kolei wartość początkowa netto odpowiada pełnej wartości początkowej zmniejszonej o zużycie wyrażone wartościowo /amortyzacja/. Posługując się wartością początkową opieramy się faktycznie na cenach środków trwałych. Te same środki trwałe wyprodukowane w różnych okresach mają zatem różną cenę. Trudność tę można w pewnym sensie wyeliminować posługując się wartością odtworzeniową środków trwałych /brutto - wyrażoną w pełnej wysokości i netto - stanowiącą ich wartość rzeczywistą/, która stanowi wartość całkowitego kosztu odtworzenia środków trwałych o podobnych własnościach użytkowych, ustaloną przy uwzględnieniu cen i norm obowiązujących w dniu wyceny. Kwestią sporną jest sprawa wyboru między wartością brutto oraz netto środków trwałych. Wynika ona m. in. z faktu, że amortyzacja księgowa dokonywana jest w sposób formalny, nie oddając rzeczywistego zużycia tych środków.

Wykorzystanie do badań każdego z osobna miernika uprzemysłowienia prowadzi często do znacznych rozbieżności w ocenach końcowych.

W niniejszej pracy /patrz podrozdział 6.2/ do oceny uprzemysłowienia /zarówno poziomu jak i profilu/ województw przyjęto jako cechę liczbę zatrudnionych /przy badaniu jednocechowym/.

Przystępując do badania wielocechowego zadanie nasze polega na ustaleniu zestawu cech diagnostycznych, które w sposób możliwie pełny charakteryzują uprzemysłowienie województw Polski w przekroju gałęziowym. Każdy miernik rozpatrywany oddzielnie nie może stanowić dostatecznej podstawy do oceny przestrzennego zróżnicowania województw ze względu na gałęziowy poziom oraz profil uprzemysłowienia.

Biorąc pod uwagę bogaty dorobek badawczy z zakresu uprzemysłowienia [86; 87; 101; 102; 121] oraz mając na uwadze dostępność

danych statystycznych, zestaw mierników uprzemysłowienia przedstawia się następująco:

1. Liczba zatrudnionych w przemyśle na 1000 mieszkańców.
2. Liczba zatrudnionych w przemyśle na 100 km² powierzchni.
3. Wartość produkcji globalnej przemysłu w mln zł na 1000 mieszkańców /w cenach porównywalnych z dnia 1.01. 1979 r./.
4. Wartość produkcji globalnej przemysłu w mln zł na 100 km² powierzchni /w cenach porównywalnych z dnia 1.01. 1979 r./.
5. Wartość brutto środków trwałych przemysłu w mln zł na 1000 mieszkańców /w cenach bieżących/.
6. Wartość brutto środków trwałych przemysłu w mln zł na 100 km² powierzchni /w cenach bieżących/.

Przyjęte zostały więc w zasadzie trzy podstawowe mierniki uprzemysłowienia /liczba zatrudnionych, wartość produkcji globalnej, wartość środków trwałych/, które odniesione zostały do liczby mieszkańców oraz powierzchni danego województwa. Cecha pierwsza nazywana jest wskaźnikiem struktury społeczno-gospodarczej, druga zaś - wskaźnikiem koncentracji przestrzennej przemysłu. Cechy trzecia i czwarta wyrażają rangę przemysłu w ekonomice kraju /województwa/. Cecha piąta wyraża stopień społecznego wyposażenia ludności w przemysłowe środki trwałe, cecha szósta zaś określa stopień koncentracji przestrzennej majątku trwałego przemysłu [87; 101].

Powyższe mierniki uprzemysłowienia stosowane będą przy analizie poziomu uprzemysłowienia województw. Liczba mieszkańców województwa oraz jego powierzchnia nie mają wpływu na proporcje międzygałęziowe w przemyśle. Zatem badając profil /typ/ uprzemysłowienia województw stosowane będą cechy:

1. Liczba zatrudnionych w przemyśle.
2. Wartość produkcji globalnej przemysłu w mln zł.
3. Wartość brutto środków trwałych przemysłu w mln zł.

Merytoryczny dobór cech diagnostycznych miał na celu wybór takich mierników uprzemysłowienia, które ujmowałyby syntetycznie zarówno czynniki /liczba zatrudnionych- czynnik ludzki, wartość środków trwałych - czynnik materialny/ jak i efekty działalności przemysłowej /mierzone w tym wypadku produkcją globalną/.

Informacje statystyczne odnośnie przyjętych w badaniu cech pochodzą z roku 1975 i 1980. Wynika to m. in. z faktu, że zaobserwowanie zmian w klasyfikacji województw ze względu na gałęziowy poziom oraz profil uprzemysłowienia wymaga pewnego okresu. Takim okresem może być okres pięcioletni. Struktury gałęziowe charakteryzują się powolnością zmian w czasie. Innym czynnikiem, który zdecydował o wyborze tych lat do badania była dostępność informacji statystycznych, które w tym zakresie opracowywane są co 5 lat, w ramach tzw. skróconego spisu przemysłowego.

Podstawowym źródłem danych statystycznych była publikacja GUS pt. "Struktura branżowa i przestrzenna przemysłu 1975 i 1980". Dostępność danych statystycznych w formie jednej publikacji GUS oraz bardzo duża obszerność materiału statystycznego zdecydowała o tym, że zrezygnowano z ich zamieszczenia w niniejszej pracy.

W tym miejscu przedstawiony zostanie ogólny schemat postępowania, który będzie wykorzystany w pracy przy badaniu struktur gałęziowych przemysłu województw Polski w roku 1975 i 1980:

I. Z zakresu struktur gałęziowych przemysłu zostaną poruszone dwa zagadnienia, tzn. klasyfikacja województw Polski ze względu na poziom oraz profil uprzemysłowienia. W związku z tym odmiennego potraktowania będzie wymagał problem normalizacji cech strukturalnych. Zatem badając poziom uprzemysłowienia /zarówno w przypadku badania jedno- jak i wielocechowego/ będziemy oceniać podobieństwo wielkości /skali/ struktur, natomiast profil uprzemysłowienia ocenimy analizując podobieństwo kształtu /formy/ struktur.

Dla prezentacji wielkości /skali/ oraz kształtu /formy/ struktur będą stosowane odpowiednio przekształcenia /2.16/ i /2.15/

przedstawione w podrozdziale 2.2.

II. W tym etapie należy zdecydować się na wybór miary podobieństwa struktur, zdając sobie sprawę, że jest to wybór w pewnym sensie arbitralny. Do tego celu autor proponuje miarę o postaci /3.16/ zaprezentowaną w podrozdziale 3.3. Wybierając tę miarę autor kierował się m. in. jej uniwersalnością /można ją stosować zarówno do badania podobieństwa wielkości jak i kształtu struktur/ oraz tym, że przyjmuje wartości pośrednie między miarą /3.18/ a /3.17/.

III. Kolejny etap badania stanowi wybór jednej metody klasyfikacji spośród bardzo wielu zaprezentowanych w rozdziale czwartym. W pierwszym rzędzie należało zdecydować się czy wybrana metoda będzie hierarchiczna czy niehierarchiczna. Wybór padł na metodę hierarchiczną z uwagi na to, że w ich wyniku otrzymuje się ciąg podziałów, od sytuacji w której każdy obiekt stanowi początkowo odrębną klasę, aż do momentu połączenia wszystkich obiektów w jedną klasę. Istnieje zatem możliwość kontrolowania procesu klasyfikacji oraz prezentacji graficznej w formie dendrogramu. Spośród metod hierarchicznych wybrano - z uwagi na jej zalety - metodę średniego połączenia międzyklasowego zaproponowaną przez M. R. Anderberga [4].

IV. Wybranie metody hierarchicznej stwarza konieczność przyjęcia reguły stop w celu ustalenia optymalnego podziału wynikowego. Do tego celu wykorzystany zostanie sposób pierwszy R. Mojeny. Z ciągu klasyfikacji - zgodnie z tym sposobem - wybiera się tę, dla której odpowiadający jej krok e / $e = 1, 2, \dots, 47$ / jako pierwszy spełnia nierówność:

$$L_{e+1} > \bar{L} + \frac{1}{2} s_L, \quad /6.1/$$

gdzie: oznaczenia takie jak we wzorze /4.9/.

Ponadto - w celu ułatwienia interpretacji otrzymanych wyników - dla każdej gałęzi w wydzielonej klasie wyznaczono średnie arytmetyczne ze znormalizowanych wartości zmiennych /jednej zmiennej w przypadku badania jednocechowego/ w obiektach ją tworzących.

Wyznaczono również uporządkowania liniowe województw według ogólnego poziomu uprzemysłowienia, które posłuży do analizy otrzymanych klas w uporządkowaniach nieliniowych. Wszystkie badane cechy są stymulantami¹, zatem chcąc ustalić wyniki liniowego uporządkowania województw według ogólnego poziomu uprzemysłowienia można posłużyć się wzorem:

$$d_r = \frac{1}{m} \sum_{l=1}^m \sum_{i=1}^k p_{ilr} \quad , \quad /6.2/$$

gdzie: d_r - wartość miary poziomu uprzemysłowienia obiektu r ,

p_{ilr} - znormalizowane według formuły /2.16/ wartości i -tego elementu struktury zmierzone dla l -tej cechy w r -tym obiekcie,

pozostałe oznaczenia jak we wzorze /2.5/.

Granice przedziałów klasyfikacyjnych wyznaczono wykorzystując konwencjonalne kryterium średniej arytmetycznej i odchylenia standardowego /por. tab. 6.1/. W przypadku badania jednocechowego przy wyznaczaniu średniej arytmetycznej oraz odchylenia standardowego pominięto województwo katowickie, które zdecydowanie góruje pod względem poziomu uprzemysłowienia nad pozostałymi województwami. Uwzględnienie w obliczeniach województwa katowickiego spowodowałoby zniekształcenie wyników obliczeń.

V. Klasyfikacja województw Polski ze względu na gałęziowy poziom oraz profil uprzemysłowienia /zarówno przy badaniu jedno- jak i wielocechowym/ będzie przeprowadzona dla roku 1975 i 1980. Pomiaru zmian w czasie jakie zaszły w otrzymanych podziałach wynikowych dokonamy w oparciu o formułę /5.21/ zaproponowaną przez W. M. Randa.

¹Określenie stymulanty i destymulanty podaje autor pracy [40] .

Tabela 6.1

Granice przedziałów klasyfikacyjnych

Granice przedziałów	Poziom uprzemysłowienia
poniżej $\bar{d}_r - \frac{1}{2} S/d_r/$	bardzo niski i niski
$(\bar{d}_r - \frac{1}{2} S/d_r/ ; \bar{d}_r >$	średni
$(\bar{d}_r ; \bar{d}_r + \frac{1}{2} S/d_r/ >$	znaczny
powyżej $\bar{d}_r + \frac{1}{2} S/d_r/$	wysoki i bardzo wysoki

\bar{d}_r - średnia arytmetyczna z wartości $d_r / r = 1, 2, \dots, n/$,

$S/d_r/$ - odchylenie standardowe.

Źródło: opracowanie własne.

6.2. Klasyfikacja województw Polski ze względu na strukturę gałęziową zatrudnienia w przemyśle.

6.2.1. Podobieństwo wielkości /skali/.

Rezultaty klasyfikacji hierarchicznej województw Polski ze względu na podobieństwo skali w strukturze gałęziowej zatrudnienia w przemyśle w roku 1975 ilustruje tabela 6.2. Z ciągu klasyfikacji - zgodnie ze sposobem pierwszym R. Mojeny - wybiera się tę, dla której odpowiadający jej krok $e / e = 1, 2, \dots, 47/$ jako pierwszy spełnia nierówność:

$$L_{e+1} > 0,3584 .$$

Zatem proces klasyfikacji należy przerwać po 38. iteracji, otrzymując ostatecznie podział zbioru województw na następujące klasy /w kolejności według liczebności/:

I. województwa: ciechanowskie, elbląskie, gorzowskie, koszalińskie, krośnieńskie, leszczyńskie, nowosądeckie, olsztyńskie, ostrołęckie, pilskie, płockie, przemyskie, siedleckie, słupeckie, suwalskie, wrocławskie, zamojskie;

II. województwa: bielskie, bydgoskie, gdańskie, kieleckie, miejskie krakowskie, lubelskie, opolskie, poznańskie, radomskie, rzeszowskie, szczecińskie, wrocławskie;

III. województwa: białostockie, częstochowskie, jeleniogórskie, kaliskie, piotrkowskie, wałbrzyskie, zielonogórskie;

IV. województwa: białkopodlaskie, chełmskie, łomżyńskie;

V. województwa: tarnobrzeskie, tarnowskie, toruńskie;

VI. województwa: sieradzkie, skierniewickie;

VII. województwo stołeczne warszawskie;

VIII. województwo katowickie;

IX. województwo konińskie;

X. województwo legnickie;

XI. województwo miejskie łódzkie.

W wyodrębnionych klasach znajdują się województwa o zbliżonym gałęziowym poziomie uprzemysłowienia. Oczywiście na otrzymane wyniki należy spojrzeć z pewną rezerwą, jako że przeprowadzono badanie jednocechowe, w którym miernikiem uprzemysłowienia jest zatrudnienie. O ograniczoności badań jednocechowych była mowa w podrozdziale 6.1. W dalszej części tego rozdziału przedstawione zostaną również wyniki badań wielocechowych w tym zakresie.

Tabela 6.2

Rezultaty klasyfikacji województw Polski ze względu na podobieństwo skali
w strukturze galeźzowej zatrudnienia / 1975 r./

Nr iteracji	Nry połączonych klas	Odległość międzyklasowa	Nr iteracji	Nry połączonych klas	Odległość międzyklasowa
1	32 - 47	0,0662	25	2 - 23	0,2327
2	10 - 32	0,0759	26	9 - 19	0,2410
3	22 - 41	0,1116	27	4 - 10	0,2436
4	9 - 36	0,1139	28	7 - 33	0,2454
5	21 - 46	0,1252	29	9 - 11	0,2554
6	34 - 35	0,1344	30	12 - 45	0,2636
7	25 - 39	0,1374	31	3 - 8	0,2679
8	30 - 49	0,1428	32	9 - 31	0,2787
9	33 - 48	0,1473	33	7 - 28	0,2809
10	17 - 21	0,1536	34	4 - 5	0,2965
11	22 - 34	0,1540	35	3 - 12	0,3106
12	10 - 15	0,1543	36	4 - 18	0,3309
13	9 - 25	0,1582	37	7 - 9	0,3313
14	43 - 44	0,1604	38	2 - 6	0,3577
15	5 - 27	0,1741	39	3 - 4	0,4117
16	37 - 38	0,1752	40	7 - 37	0,4128
17	11 - 26	0,1818	41	3 - 42	0,4263
18	33 - 40	0,1832	42	7 - 16	0,4528
19	13 - 30	0,1927	43	2 - 7	0,4848
20	42 - 43	0,2005	44	3 - 20	0,5441
21	19 - 29	0,2007	45	1 - 24	0,5897
22	7 - 17	0,2121	46	2 - 3	0,6014
23	5 - 22	0,2221	47	1 - 14	0,7426
24	3 - 13	0,2288	48	1 - 2	0,8212

Źródło: obliczenia własne.

Rys. 6.1. Klasyfikacja województw Polski ze względu na podobieństwo skali w strukturze gałęziowej zatrudnienia w 1975 r.



- klasa I
- klasa II
- klasa III
- klasa IV



- klasa V
- klasa VI
- klasy jednoelementowe

Tabela 6.3

Średnie arytmetyczne ze znormalizowanych /wg formuły /2.46// wartości zmiennej zatrudnienia dla każdej gałęzi w wydzielonej klasie /1975 r./

Nr klasy \ Nr gałęzi	1	2	3	4	5	6	7	8	9
1	0,0049	0,0001	0,0335	0,0062	0,0090	0,0138	0,0168	0,0226	0,0035
2	0,0117	0,0151	0,1667	0,0267	0,0271	0,0235	0,0473	0,0479	0,0104
3	0,0165	0,0080	0,0738	0,0121	0,0291	0,0208	0,0836	0,0231	0,0064
4	0,0004	0,0000	0,0087	0,0003	0,0079	0,0038	0,0178	0,0102	0,0014
5	0,0059	0,0100	0,0677	0,0437	0,0151	0,0084	0,0171	0,0277	0,0031
6	0,0010	0,0000	0,0233	0,0072	0,0107	0,0030	0,0460	0,0141	0,0012
7	0,0201	0,0284	0,5368	0,0654	0,0435	0,0326	0,0722	0,0712	0,0616
8	1,0000	0,3140	0,5126	0,1003	0,1250	0,0328	0,1006	0,0940	0,0197
9	0,0310	0,0073	0,0190	0,0051	0,0034	0,0029	0,0155	0,0170	0,0032
10	0,0035	0,0744	0,0444	0,0030	0,0118	0,0082	0,0296	0,0185	0,0049
11	0,0145	0,0002	0,1719	0,0570	0,0165	0,0213	0,4600	0,0402	0,0201

Źródło: obliczenia własne.

Wartości miary poziomu uprzemysłowienia /1975 r./

Lp.	Województwo	Wartości miary
1	katowickie	2,2990
2	stołeczne warszawskie	0,9318
3	miejskie łódzkie	0,8017
4	gdańskie	0,4644
5	wrocławskie	0,4617
6	bielskie	0,4507
7	poznańskie	0,4432
8	wałbrzyskie	0,4384
9	miejskie krakowskie	0,4346
10	kieleckie	0,4223
11	opolskie	0,4138
12	bydgoskie	0,3966
13	częstochowskie	0,3163
14	suczocińskie	0,2980
15	jeleniogórskie	0,2824
16	lubelskie	0,2698
17	zielonogórskie	0,2432
18	plotrkowskie	0,2415
19	radomskie	0,2397
20	kaliszkie	0,2396
21	toruńskie	0,2177
22	rzeszowskie	0,2095
23	legnickie	0,1983
24	tarnobrzeskie	0,1902
25	tarnowskie	0,1890
26	białostockie	0,1736
27	olsztyńskie	0,1640
28	gorzowskie	0,1537
29	krośnieńskie	0,1536
30	płockie	0,1384
31	nowosądeckie	0,1359
32	skierniewickie	0,1188
33	olbąskie	0,1181
34	pilskie	0,1176
35	koszalińskie	0,1152
36	siódleckie	0,1113
37	ślupskie	0,1081
38	konińskie	0,1044
39	włocławskie	0,1005
40	sioradzkie	0,0945
41	suwalskie	0,0843
42	łoszczyńskie	0,0841
43	zamojskie	0,0825
44	przemyskie	0,0821
45	ostrołęckie	0,0676
46	ciechanowskie	0,0662
47	chełmskie	0,0587
48	łomżyńskie	0,0473
49	białskopodlaskie	0,0461

Poziom uprzemysłowienia :

bardzo niski i niski

poniżej 0,1438

średni

(0,1438 ; 0,2359)

znaczący

(0,2359 ; 0,3280)

wysoki i bardzo wysoki

powyżej 0,3280

Źródło: obliczenia własne.

Graficzne wyniki klasyfikacji województw Polski ze względu na podobieństwo skali w strukturze gałęziowej zatrudnienia w 1975 roku ilustruje rys. 6.1. Ponadto - w celu ułatwienia interpretacji otrzymanych wyników - dla każdej gałęzi w wydzielonej klasie wyznaczono średnie arytmetyczne ze znormalizowanych wartości zmiennej zatrudnienie i zaprezentowano je w tabeli 6.3. Wyznaczono również uporządkowanie liniowe województw według ogólnego poziomu uprzemysłowienia /por. tab. 6.4/, które posłuży do analizy otrzymanych klas.

Najwyższy poziom uprzemysłowienia prezentuje klasa ósma utworzona przez województwo katowickie. Przewodzi ona - pod względem uprzemysłowienia - prawie we wszystkich gałęziach przemysłu /oprócz przemysłu lekkiego i pozostałych gałęzi przemysłu/. O poziomie uprzemysłowienia tej klasy decydują przede wszystkim przemysły: paliwowo-energetyczny, elektromaszynowy i metalurgiczny.

O bardzo wysokim poziomie uprzemysłowienia województwa stołecznego warszawskiego i miejskiego łódzkiego /tworzących klasy jednoelementowe o numerach siedem i jedenaście/ decyduje głównie przemysł odpowiednio elektromaszynowy i lekki. W wymienionych gałęziach klasy te prezentują najwyższy - wśród wydzielonych klas - poziom uprzemysłowienia.

Najniższy poziom uprzemysłowienia przedstawia klasa czwarta obejmująca województwa: białkopodlaskie, chełmskie i łomżyńskie. Ponadto cechą charakterystyczną tej klasy jest bardzo niski poziom uprzemysłowienia we wszystkich gałęziach przemysłu. O poziomie uprzemysłowienia tej klasy decyduje w znacznej mierze przemysł lekki. Nie występuje w niej gałąź przemysłu metalurgicznego.

Różnice między klasami szóstą oraz dziewiątą, które obejmują województwa o niskim poziomie uprzemysłowienia wynikają z roli jaką odgrywają w nich poszczególne gałęzie przemysłu. W klasie szóstej na ogólny poziom uprzemysłowienia największy wpływ ma przemysł lekki, w klasie dziewiątej zaś - przemysł paliwowo-ener-

getyczny.

Zdecydowana większość województw klasy pierwszej prezentuje niski poziom uprzemysłowienia /za wyjątkiem województwa olsztyńskiego, gorzowskiego i krośnieńskiego, które przedstawiają średni poziom uprzemysłowienia/. O poziomie uprzemysłowienia województw tej klasy decyduje przede wszystkim przemysł elektromaszynowy i spożywczy.

Klasy piąta i dziesiąta przedstawiają średni poziom uprzemysłowienia. O poziomie uprzemysłowienia tych klas decydują odpowiednio przemysł elektromaszynowy i chemiczny oraz metalurgiczny i elektromaszynowy.

W skład klasy trzeciej wchodzi województwa o wysokim /wałbrzyskie/, średnim /białostockie/ i znacznym poziomie uprzemysłowienia /pozostałe województwa tej klasy/. W klasie tej na ogólny poziom uprzemysłowienia największy wpływ ma przemysł lekki i elektromaszynowy.

Klasa druga skupia województwa o wysokim i znacznym poziomie uprzemysłowienia. Jedynie województwo rzeszowskie prezentuje średni poziom uprzemysłowienia. Oprócz Bielska-Białej i Radomia pozostałe stolice województw tej klasy były stolicami województw w starym podziale administracyjnym. W klasie tej na ogólny poziom uprzemysłowienia znaczny wpływ ma przemysł elektromaszynowy, dużo niższy zaś, ale wyraźniejszy od innych gałęzi, ma również przemysł lekki oraz spożywczy.

Województwa znajdujące się w jednej klasie klasyfikacji na ogół zajmują zbliżone pozycje po uszeregowaniu ich według wartości poziomu uprzemysłowienia.

Rezultaty klasyfikacji hierarchicznej województw Polski ze względu na podobieństwo skali w strukturze gałęziowej zatrudnienia w roku 1980 ilustruje tabela 6.5. Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok e / $e = 1, 2, \dots, 47$ /

jako pierwszy spełnia nierówność:

$$L_{e+1} > 0,3590 .$$

Przeto proces klasyfikacji należy przerwać po 38. iteracji, otrzymując ostatecznie podział zbioru województw na następujące klasy:

I. województwa: białostockie, elbląskie, gorzowskie, koszalińskie, krośnieńskie, nowosądeckie, olsztyńskie, pilskie, płockie, siedleckie, sieradzkie, skierniewickie, słupeckie;

II. województwa: bielskie, bydgoskie, gdańskie, kieleckie, miejskie krakowskie, lubelskie, opolskie, poznańskie, radomskie, rzeszowskie, szczecińskie, wrocławskie;

III. województwa: białkopodlaskie, ciechanowskie, leszczyńskie, łomżyńskie, ostrołęckie, przemyskie, suwalskie, włocławskie, zamojskie;

IV. województwa: częstochowskie, jeleniogórskie, kaliskie, piotrkowskie, wałbrzyskie, zielonogórskie;

V. województwa: tarnobrzeskie, tarnowskie, toruńskie;

VI. województwo stołeczne warszawskie;

VII. województwo chełmskie;

VIII. województwo katowickie;

IX. województwo konińskie;

X. województwo legnickie;

XI. województwo miejskie łódzkie.

Wyodrębnione klasy zawierają województwa o zbliżonym gałęziowym poziomie uprzemysłowienia.

Graficzne wyniki klasyfikacji województw Polski ze względu na podobieństwo skali w strukturze gałęziowej zatrudnienia w 1980 roku prezentuje rys. 6.2. Ponadto - w celu ułatwienia interpretacji

Tabela 6.5
 Rezultaty klasyfikacji województw Polski ze względu na podobieństwo skali
 w strukturze gałęziowej zatrudnienia /1980 r./

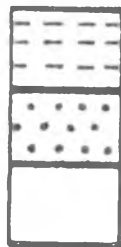
Nr iteracji	Nry połączonych klas	Odległość międzyklasowa	Nr iteracji	Nry połączonych klas	Odległość międzyklasowa
1	32 - 47	0,0578	25	4 - 10	0,2243
2	10 - 32	0,0662	26	3 - 29	0,2341
3	7 - 21	0,0834	27	11 - 19	0,2434
4	22 - 35	0,0913	28	8 - 12	0,2446
5	9 - 25	0,1002	29	7 - 33	0,2510
6	9 - 39	0,1116	30	2 - 23	0,2606
7	22 - 34	0,1127	31	3 - 11	0,2622
8	9 - 36	0,1404	32	4 - 5	0,2659
9	12 - 30	0,1455	33	3 - 31	0,2803
10	22 - 41	0,1550	34	7 - 28	0,2887
11	33 - 48	0,1571	35	4 - 18	0,2930
12	10 - 15	0,1574	36	8 - 45	0,3033
13	7 - 46	0,1600	37	2 - 7	0,3407
14	5 - 27	0,1603	38	3 - 37	0,3423
15	43 - 44	0,1623	39	3 - 42	0,3737
16	13 - 49	0,1635	40	4 - 8	0,3753
17	9 - 17	0,1668	41	2 - 3	0,4183
18	37 - 38	0,1719	42	2 - 16	0,4611
19	3 - 26	0,1775	43	2 - 6	0,5491
20	33 - 40	0,1931	44	4 - 20	0,5532
21	8 - 13	0,1962	45	1 - 4	0,5794
22	42 - 43	0,1972	46	1 - 2	0,6041
23	5 - 22	0,2034	47	1 - 24	0,7536
24	3 - 9	0,2124	48	1 - 14	0,9186

Źródło: obliczenia własne.

Rys. 6.2. Klasyfikacja województw Polski ze względu na podobieństwo skali w strukturze gałęziowej zatrudnienia w 1980 r.



- klasa I
- klasa II
- klasa III



- klasa IV
- klasa V
- klasy jednoelementowe

Tabela 6.6

Średnie arytmetyczne ze znormalizowanych /wg formuły /2.16// wartości zmiennej zatrudnienia dla każdej gałęzi w wydzielonej klasie /1980 r./

Nr klasy	Nr gałęzi	1	2	3	4	5	6	7	8	9
1		0,0053	0,0000	0,0409	0,0082	0,0091	0,0132	0,0246	0,0205	0,0054
2		0,0108	0,0140	0,1549	0,0235	0,0235	0,0200	0,0409	0,0419	0,0137
3		0,0017	0,0000	0,0207	0,0024	0,0060	0,0089	0,0114	0,0170	0,0039
4		0,0191	0,0064	0,0765	0,0111	0,0276	0,0205	0,0739	0,0200	0,0103
5		0,0067	0,0092	0,0647	0,0419	0,0132	0,0083	0,0163	0,0255	0,0047
6		0,0188	0,0258	0,4753	0,0523	0,0429	0,0246	0,0552	0,0535	0,0603
7		0,0006	0,0000	0,0056	0,0002	0,0156	0,0027	0,0162	0,0095	0,0012
8		1,0000	0,3028	0,5132	0,0841	0,1200	0,0262	0,0846	0,0793	0,0287
9		0,0328	0,0061	0,0207	0,0050	0,0043	0,0029	0,0219	0,0156	0,0076
10		0,0036	0,0802	0,0457	0,0027	0,0099	0,0073	0,0270	0,0157	0,0062
11		0,0137	0,0002	0,1536	0,0448	0,0125	0,0159	0,3650	0,0322	0,0210

Źródło: obliczenia własne.

Wartości miary poziomu uprzemysłowienia /1980 r./

Lp.	Województwo	Wartości miary
1	katowickie	2,2389
2	stołeczne warszawskie	0,8087
3	miejskie łódzkie	0,6589
4	gdańskie	0,4130
5	wrocławskie	0,4055
6	poznańskie	0,3976
7	bielskie	0,3974
8	miejskie krakowskie	0,3960
9	kieleckie	0,3869
10	wałbrzyskie	0,3858
11	opolskie	0,3796
12	bydgoskie	0,3662
13	częstochoowskie	0,2870
14	szczecińskie	0,2808
15	lubelskie	0,2579
16	jeleniogórskie	0,2488
17	radomskie	0,2265
18	zielenogórskie	0,2262
19	kaliskie	0,2241
20	piotrkowskie	0,2232
21	rzeszowskie	0,2173
22	toruńskie	0,2001
23	łognickie	0,1983
24	tarnobrzeskie	0,1895
25	tarnowskie	0,1827
26	olsztyńskie	0,1669
27	białostockie	0,1663
28	krośnieńskie	0,1552
29	gorzowskie	0,1435
30	płockie	0,1400
31	nowosądeckie	0,1305
32	konińskie	0,1169
33	elbląskie	0,1156
34	pilskie	0,1137
35	siedleckie	0,1111
36	koszalińskie	0,1097
37	ślupskie	0,1062
38	skierniewickie	0,1053
39	włocławskie	0,0968
40	sieradzkie	0,0922
41	przemyskie	0,0816
42	leszczyńskie	0,0811
43	suwalskie	0,0800
44	ciechanowskie	0,0733
45	zamojskie	0,0725
46	ostrolęckie	0,0680
47	chełmskie	0,0516
48	łomżyńskie	0,0500
49	białkopodlaskie	0,0470

Poziom uprzemysłowienia:

bardzo niski i niski

poniżej 0,1391

średni

(0,1391 ; 0,2174 >

znaczny

(0,2174 ; 0,2957 >

wysoki i bardzo wysoki

powyżej 0,2957

Źródło: obliczenia własne.

otrzymanych wyników - dla każdej gałęzi w wydzielonej klasie wyznaczono średnie arytmetyczne ze znormalizowanych wartości zmiennej zatrudnienie i zaprezentowano je w tabeli 6.6. Wyznaczono również uporządkowanie liniowe województw według ogólnego poziomu uprzemysłowienia /por. tab. 6.7/, które posłuży do analizy wydzielonych klas.

Najwyższy poziom uprzemysłowienia przedstawia województwo katowickie tworzące klasę ósmą. Przeważa ono pod względem poziomu uprzemysłowienia prawie we wszystkich gałęziach przemysłu /oprócz przemysłu lekkiego i pozostałych gałęzi przemysłu/. W województwie tym o poziomie uprzemysłowienia decydują przede wszystkim przemysły: paliwowo-energetyczny, elektromaszynowy i metalurgiczny.

Województwo stołeczne warszawskie oraz miejskie łódzkie tworzące klasy jednoelementowe /szóstą i jedenastą/ prezentują bardzo wysoki poziom uprzemysłowienia, na który decydujący wpływ ma przemysł odpowiednio elektromaszynowy i lekki. W wymienionych gałęziach klasy te prezentują najwyższy - wśród wydzielonych klas - poziom uprzemysłowienia.

Cechą charakterystyczną powyższych trzech klas jest to, że skupiają województwa, których stolice są największymi aglomeracjami miejskimi w Polsce.

Bardzo niski poziom uprzemysłowienia - i to we wszystkich gałęziach przemysłu - prezentują klasy trzecia i siódma. W klasie trzeciej spośród innych gałęzi wyróżnia się przemysł elektromaszynowy i spożywczy, natomiast w klasie siódmej - lekki i mineralny. Nie występuje w tych klasach gałąź przemysłu metalurgicznego.

Klasa dziewiąta przedstawia niski poziom uprzemysłowienia, na który znaczący wpływ ma przemysł paliwowo-energetyczny.

Klasa pierwsza utworzona jest z województw o niskim i średnim poziomie uprzemysłowienia. Znaczący wpływ na ogólny poziom uprzemysłowienia województw tej klasy ma przemysł elektromaszynowy. W klasie tej nie występuje przemysł metalurgiczny.

Różnice między klasami piątą i dziesiątą, które prezentują średni poziom uprzemysłowienia, wynikają z roli jaką odgrywają w nich poszczególne gałęzie przemysłu. W klasie piątej największy wpływ na ogólny poziom uprzemysłowienia ma przemysł elektromaszynowy i chemiczny, w klasie dziesiątej zaś - przemysł metalurgiczny i elektromaszynowy.

W skład klasy czwartej wchodzi województwa o wysokim /wałbrzyskie/ i znacznym ogólnym poziomie uprzemysłowienia. W klasie tej spośród innych gałęzi wyróżnia się przemysł elektromaszynowy i lekki.

Klasa druga skupia województwa o wysokim i znacznym poziomie uprzemysłowienia. Jedynie województwo rzeszowskie prezentuje średni poziom uprzemysłowienia. Oprócz Bielska-Białej i Radomia pozostałe stolice województw tej klasy były stolicami województw w starym podziale administracyjnym. W klasie tej o poziomie uprzemysłowienia decyduje w znacznej mierze przemysł elektromaszynowy. Dużo niższy wpływ, ale wyraźniejszy od innych gałęzi, ma przemysł lekki oraz spożywczy.

Województwa znajdujące się w jednej klasie klasyfikacji na ogół zajmują zbliżone pozycje po uszeregowaniu ich według wartości poziomu uprzemysłowienia.

6.2.2. Podobieństwo kształtu /formy/..

Rezultaty klasyfikacji hierarchicznej województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej zatrudnienia w roku 1975 przedstawia tabela 6.8. Z ciągu klasyfikacji - zgodnie z przyjętym kryterium - wybiera się tę, dla której odpowiadający jej krok e / $e = 1, 2, \dots, 47$ / jako pierwszy spełnia nierówność:

$$L_{e+1} > 0,2394 .$$

Proces klasyfikacji przerywa się więc po 37. iteracji, otrzymując w efekcie podział zbioru województw na następujące klasy:

I. województwa: stołeczne warszawskie, bielskie, bydgoskie, elbląskie, gdańskie, kaliskie, kielecki, krośnieńskie, lubelskie, nowosądeckie, opolskie, płockie, poznańskie, radomskie, rzeszowskie, siedleckie, szczecińskie, wrocławskie;

II. województwa: białkopodlaskie, ciechanowskie, koszalińskie, leszczyńskie, olsztyńskie, przemyskie, śląskie, suwalskie, wrocławskie, zamojskie;

III. województwa: białostockie, jeleniogórskie, łomżyńskie, miejskie łódzkie, piotrkowskie, sieradzkie, skierniewickie, wałbrzyskie, zielonogórskie;

IV. województwa: miejskie krakowskie, tarnobrzeskie;

V. województwa: ostrołęckie, pilskie;

VI. województwa: tarnowskie, toruńskie;

VII. województwo chełmskie;

VIII. województwo częstochowskie;

IX. województwo gorzowskie;

X. województwo katowickie;

XI. województwo konińskie;

XII. województwo legnickie.

Wyodrębnione klasy zawierają województwa o zbliżonym gałęziowym profilu uprzemysłowienia.

Graficzne wyniki klasyfikacji województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej zatrudnienia w roku 1975 przedstawia rys. 6.3. Ponadto - w celu ułatwienia interpretacji otrzymanych wyników - dla każdej gałęzi w wydzielonej klasie policzono średnie arytmetyczne ze znormalizowanych wartości zmien-

Tabela 6.8

Rezultaty klasyfikacji województw Polski ze względu na podobieństwo kształtu
w strukturze gałęziowej zatrudnienia /1975 r./

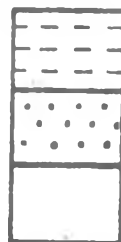
Nr iteracji	Nry połączonych klas	Odległość międzyklasowa	Nr, iteracji	Nry połączonych klas	Odległość międzyklasowa
1	10 - 32	0,0497	25	23 - 24	0,1617
2	1 - 47	0,0564	26	40 - 48	0,1726
3	10 - 22	0,0610	27	4 - 9	0,1738
4	1 - 35	0,0644	28	28 - 29	0,1751
5	7 - 21	0,0660	29	2 - 40	0,1864
6	2 - 39	0,0728	30	12 - 45	0,1878
7	1 - 10	0,0796	31	18 - 42	0,1972
8	9 - 13	0,0800	32	3 - 12	0,1982
9	3 - 49	0,0874	33	1 - 31	0,2070
10	9 - 36	0,1026	34	2 - 7	0,2139
11	7 - 17	0,1036	35	4 - 19	0,2196
12	25 - 27	0,1127	36	1 - 4	0,2264
13	5 - 41	0,1181	37	3 - 23	0,2393
14	2 - 33	0,1220	38	14 - 16	0,2474
15	1 - 34	0,1259	39	11 - 43	0,2491
16	23 - 37	0,1288	40	2 - 28	0,2684
17	7 - 46	0,1298	41	8 - 18	0,2839
18	3 - 30	0,1386	42	1 - 8	0,3035
19	9 - 25	0,1429	43	1 - 11	0,3165
20	43 - 44	0,1438	44	1 - 2	0,3204
21	1 - 15	0,1460	45	3 - 6	0,3294
22	2 - 26	0,1487	46	1 - 3	0,3997
23	3 - 38	0,1573	47	1 - 20	0,5142
24	1 - 5	0,1574	48	1 - 14	0,5349

Źródło: obliczenia własne.

Rys. 6.3. Klasyfikacja województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej zatrudnienia w 1975 r.



- klasa I
- klasa II
- klasa III
- klasa IV



- klasa V
- klasa VI
- klasy jednoelementowe

Tabela 6.9

Średnie arytmetyczne ze znormalizowanych /wg formuły /2.15// wartości zmiennej zatrudnienie dla każdej gałęzi w wydzielonej klasie /1975 r./

Nr klasy	Nr gałęzi	Nr gałęzi								
		1	2	3	4	5	6	7	8	9
1	a	0,1038	0,0356	1,0000	0,1565	0,1735	0,1663	0,3526	0,3371	0,0643
	b	4,34	1,49	41,85	6,55	7,26	6,96	14,75	14,11	2,69
2	a	0,0880	0,0000	0,9279	0,1060	0,2836	0,4665	0,6290	0,8866	0,1415
	b	2,50	0	26,29	3,00	8,04	13,22	17,82	25,12	4,01
3	a	0,1139	0,0020	0,6237	0,1398	0,2681	0,1904	1,0000	0,2720	0,0571
	b	4,27	0,07	23,39	5,24	10,05	7,14	37,50	10,20	2,14
4	a	0,1241	0,5455	1,0000	0,3979	0,2801	0,1764	0,2707	0,3056	0,0931
	b	3,89	17,08	31,31	12,46	8,77	5,52	8,48	9,57	2,92
5	a	0,1243	0,0020	1,0000	0,0609	0,4214	0,5992	0,1603	0,5125	0,0867
	b	4,19	0,07	33,70	2,05	14,20	20,20	5,40	17,27	2,92
6	a	0,0903	0,0458	1,0000	0,7523	0,2403	0,1584	0,2647	0,5061	0,0605
	b	2,89	1,47	32,07	24,12	7,71	5,08	8,49	16,23	1,94
7	a	0,0339	0,0000	0,2527	0,0143	0,9281	0,1541	1,0000	0,5124	0,0318
	b	1,16	0	8,94	0,49	31,60	5,25	34,04	17,44	1,08
8	a	0,1051	0,5955	1,0000	0,1260	0,2472	0,3255	0,9123	0,1945	0,1680
	b	2,86	16,21	27,22	3,43	6,73	8,86	24,83	5,29	4,57
9	a	0,1188	0,0000	1,0000	0,8319	0,1765	0,7769	0,6672	0,5938	0,0856
	b	2,79	0	23,53	19,57	4,15	18,28	15,70	13,97	2,01
10	a	1,0000	0,3140	0,5126	0,1003	0,1250	0,0328	0,1006	0,0940	0,0197
	b	43,50	13,66	22,29	4,36	5,44	1,43	4,37	4,09	0,86
11	a	1,0000	0,2357	0,6135	0,1650	0,1106	0,0937	0,5006	0,5482	0,1053
	b	29,65	6,99	18,19	4,89	3,28	2,78	14,84	16,26	3,12
12	a	0,0478	1,0000	0,5969	0,0406	0,1589	0,1103	0,3982	0,2494	0,0670
	b	1,79	37,47	22,36	1,52	5,95	4,13	14,92	9,35	2,51

a - w liczbach bezwzględnych,

b - w %.

Źródło: obliczenia własne.

nej zatrudnienie i zaprezentowano je w tabeli 6.9.

Klasa pierwsza obejmuje województwa, w których w gałęziowym profilu uprzemysłowienia zdecydowanie dominuje przemysł elektromaszynowy. Znaczną rolę w porównaniu z pozostałymi gałęziami w tej klasie odgrywają również przemysły lekki oraz spożywczy.

W klasie drugiej wyróżnić można w gałęziowym profilu uprzemysłowienia przemysł elektromaszynowy oraz spożywczy. Nieco mniejszą rolę odgrywa przemysł lekki oraz drzewno-papierniczy. W tej klasie nie jest reprezentowana gałąź przemysłu metalurgicznego.

W gałęziowym profilu uprzemysłowienia województw tworzących klasę trzecią uwidacznia się przemysł lekki. Niepoślednie znaczenie w tej klasie ma również przemysł elektromaszynowy.

W kolejnych trzech klasach pomimo, że w gałęziowym profilu uprzemysłowienia przeważa przemysł elektromaszynowy, to jednak występują między nimi zasadnicze różnice. Wynikają one z roli jaką w tych klasach odgrywają inne gałęzie przemysłu. W klasie czwartej wyróżnia się ponadto przemysł metalurgiczny i chemiczny, w klasie piątej - mineralny, spożywczy oraz drzewno-papierniczy, a w klasie szóstej - przemysł chemiczny i spożywczy.

W klasie siódmej w gałęziowym profilu uprzemysłowienia górują przemysły: lekki i mineralny. Niskie znaczenie ma przemysł paliwowo-energetyczny, chemiczny i pozostałe gałęzie przemysłu. Nie występuje w tej klasie gałąź przemysłu metalurgicznego.

Województwo częstochowskie tworzące klasę ósmą cechuje dominacja w gałęziowym profilu uprzemysłowienia przemysłu elektromaszynowego i lekkiego. Doniosłe znaczenie w tej klasie ma również przemysł metalurgiczny.

Klasę dziewiątą tworzy województwo gorzowskie, w którym przeważa w gałęziowym profilu uprzemysłowienia przemysł elektromaszynowy, a prawie równorzędne znaczenie mają przemysły: chemiczny, drzewno-papierniczy, lekki i spożywczy. W klasie tej nie jest reprezentowana gałąź przemysłu metalurgicznego.

W kolejnej klasie jednoelementowej utworzonej przez województwo katowickie wybija się w gałęziowym profilu uprzemysłowienia przemysł paliwowo-energetyczny. Znaczny wpływ ma w tym województwie również przemysł elektromaszynowy oraz metalurgiczny. Pozostałe przemysły w tej klasie nie odgrywają poważniejszej roli.

Województwo konińskie tworzące klasę jedenastą cechuje dominacja w gałęziowym profilu uprzemysłowienia przemysłu paliwowo-energetycznego. Dużą rolę odgrywają w tym województwie również przemysły: elektromaszynowy, lekki i spożywczy.

W województwie legnickim na plan pierwszy wybija się przemysł metalurgiczny. Kolejno wyróżniają się w gałęziowym profilu uprzemysłowienia przemysły elektromaszynowy i lekki.

Mianem klasy wyspecjalizowanej w danej gałęzi przemysłu będzie nazywana taka, której udział w całości wynosi co najmniej 25%

W przemyśle paliwowo-energetycznym wyspecjalizowała się klasa dziesiąta, w przemyśle metalurgicznym - klasa dwunasta, w przemyśle elektromaszynowym - klasy: pierwsza, druga, czwarta, piąta, szósta i ósma, w przemyśle mineralnym - klasa siódma, w przemyśle lekkim - klasa trzecia i siódma, w przemyśle spożywczym zaś - klasa druga. Nie wyspecjalizowała się żadna klasa w przemyśle chemicznym, drzewno-papierniczym oraz pozostałych gałęziach przemysłu.

Rezultaty klasyfikacji hierarchicznej województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej zatrudnienia w roku 1980 prezentuje tabela 6.10. Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok $e / e = 1, 2, \dots, 47$ jako pierwszy spełnia nierówność:

$$\mathcal{L}_{e+1} > 0,2398 .$$

Proces klasyfikacji przerywa się więc po 36. kroku, otrzymując w rezultacie podział zbioru województw na następujące klasy:

- I. województwa: białkopodlaskie, białostockie, bielskie, ciechanowskie, częstochowskie, elbląskie, kaliskie, koszalińskie, krośnieńskie, leszczyńskie, nowosądeckie, olsztyńskie, opolskie, siedleckie, słupskie, szczecińskie, włocławskie, zielonogórskie;
- II. województwa: stołeczne warszawskie, bydgoskie, gdańskie, kieleckie, lubelskie, pilskie, płockie, poznańskie, radomskie, rzeszowskie, wrocławskie;
- III. województwa: łomżyńskie, miejskie łódzkie, sieradzkie, skiernewickie;
- IV. województwa: jeleniogórskie, piotrkowskie, wałbrzyskie;
- V. województwa: przemyskie, suwalskie, zamojskie;
- VI. województwa: miejskie krakowskie, tarnobrzesckie;
- VII. województwa: tarnowskie, toruńskie;
- VIII. województwo chełmskie;
- IX. województwo gorzowskie;
- X. województwo katowickie;
- XI. województwo konińskie;
- XII. województwo legnickie;
- XIII. województwo ostrołęckie.

W wydzielonych klasach znajdują się województwa o zbliżonym gałęziowym profilu uprzemysłowienia.

Graficzne wyniki klasyfikacji województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej zatrudnienia w roku 1980 ilustruje rys. 6.4. Ponadto - w celu ułatwienia interpretacji otrzymanych wyników - dla każdej gałęzi w wydzielonej klasie wyznaczono średnie arytmetyczne ze znormalizowanych wartości zmiennej zatrudnienie i zaprezentowano je w tabeli 6.11.

Tabela 6.10

Rezultaty klasyfikacji województw Polski ze względu na podobieństwo kształtu
w strukturze gałęziowej zatrudnienia /1980 r./

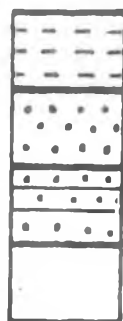
Nr iteracji	Nry połączonych klas	Odległość międzyklasowa	Nr iteracji	Nry połączonych klas	Odległość międzyklasowa
1	2 - 13	0,0454	25	33 - 48	0,1583
2	10 - 32	0,0455	26	7 - 17	0,1616
3	35 - 47	0,0487	27	1 - 5	0,1743
4	22 - 34	0,0537	28	23 - 38	0,1757
5	10 - 35	0,0592	29	18 - 42	0,1795
6	2 - 9	0,0651	30	12 - 45	0,1802
7	7 - 21	0,0719	31	33 - 40	0,1916
8	3 - 39	0,0803	32	2 - 4	0,1922
9	1 - 10	0,0868	33	1 - 31	0,1924
10	1 - 22	0,0953	34	23 - 24	0,2070
11	2 - 25	0,0990	35	2 - 19	0,2217
12	7 - 41	0,1119	36	2 - 7	0,2254
13	17 - 26	0,1145	37	1 - 2	0,2543
14	2 - 36	0,1161	38	12 - 23	0,2603
15	3 - 49	0,1247	39	18 - 43	0,2643
16	24 - 37	0,1378	40	11 - 33	0,2754
17	17 - 46	0,1389	41	1 - 18	0,2918
18	19 - 27	0,1411	42	14 - 16	0,2943
19	5 - 29	0,1451	43	11 - 28	0,3038
20	4 - 8	0,1466	44	1 - 11	0,3291
21	2 - 3	0,1498	45	6 - 12	0,3409
22	1 - 15	0,1503	46	1 - 6	0,3917
23	12 - 30	0,1532	47	1 - 14	0,5255
24	43 - 44	0,1568	48	1 - 20	0,5397

Źródło: obliczenia własne.

Rys. 6:4. Klasyfikacja województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej zatrudnienia w 1980 r.



- klasa I
- klasa II
- klasa III
- klasa IV



- klasa V
- klasa VI
- klasa VII
- klasy jednoelementowe

Tabela 6.11

Średnie arytmetyczno ze znormalizowanych /wg formuły /2,45// wartości zmiennej zatrudnionio dla każdej gałęzi w wydzielonej klasie /1980 r./

Nr klasy	Nr gałęzi	Kategorie								
		1	2	3	4	5	6	7	8	9
1	a	0,0850	0,0756	0,9964	0,1523	0,2127	0,2789	0,5342	0,4951	0,1476
	b	2,99	1,21	33,88	5,18	7,23	9,48	18,17	16,84	5,02
2	a	0,0913	0,0321	1,0000	0,1208	0,1466	0,1354	0,1967	0,2597	0,0772
	b	3,99	1,57	48,77	5,89	7,15	6,60	9,59	12,67	3,77
3	a	0,0316	0,0001	0,5560	0,1313	0,1275	0,1113	1,0000	0,3400	0,0904
	b	1,32	0	23,28	5,50	5,34	4,66	41,87	14,24	3,79
4	a	0,3521	0,0060	0,8190	0,2223	0,4550	0,2418	1,0000	0,1801	0,0781
	b	10,50	0,18	24,41	6,63	13,56	7,21	29,81	5,37	2,33
5	a	0,1028	0,0002	0,8004	0,0779	0,3334	0,5051	0,7486	0,9215	0,2399
	b	2,75	0,01	21,46	2,09	8,94	13,54	20,07	24,71	6,43
6	a	0,1380	0,4630	1,0000	0,3608	0,2556	0,0850	0,3232	0,2587	0,1177
	b	4,60	15,42	33,31	12,02	8,51	2,83	10,77	8,62	3,92
7	a	0,0896	0,0665	1,0000	0,7770	0,1995	0,1512	0,2751	0,5062	0,0913
	b	2,84	2,11	31,68	24,62	6,32	4,79	8,71	16,04	2,89
8	a	0,0404	0,0000	0,3481	0,0156	0,9653	0,1704	1,0000	0,5856	0,0783
	b	1,26	0	10,87	0,49	30,13	5,32	31,21	18,28	2,44
9	a	0,1234	0,0000	1,0000	0,7395	0,1462	0,7940	0,7183	0,5997	0,1979
	b	2,86	0	23,15	17,12	3,39	18,38	16,63	13,89	4,58
10	a	1,0000	0,3028	0,5132	0,0841	0,1200	0,0262	0,0846	0,0793	0,0287
	b	44,67	13,52	22,92	3,76	5,36	1,17	3,78	3,54	1,28
11	a	1,0000	0,1866	0,6313	0,1528	0,1325	0,0885	0,6673	0,4749	0,2321
	b	28,04	5,23	17,70	4,29	3,72	2,48	18,71	13,32	6,51
12	a	0,0456	1,0000	0,5696	0,0337	0,1240	0,0910	0,3376	0,1968	0,0780
	b	1,84	40,38	23,00	1,36	5,01	3,68	13,63	7,95	3,15
13	a	0,1886	0,0033	1,0000	0,0456	0,3421	0,8134	0,1241	0,5386	0,0888
	b	6,00	0,10	31,80	1,45	10,88	25,87	3,95	17,13	2,82

a - w liczbach bezwzględnych,

b - w %.

Źródło: obliczenia własne.

Klasa pierwsza obejmuje województwa, w których w gałęziowym profilu uprzemysłowienia przoduje przemysł elektromaszynowy. Znaczną rolę w porównaniu z innymi gałęziami tej klasy odgrywa również przemysł lekki i spożywczy.

W gałęziowym profilu uprzemysłowienia województw tworzących klasę drugą zdecydowanie góruje przemysł elektromaszynowy. Inne gałęzie w tej klasie nie odgrywają poważniejszej roli.

Województwa klasy trzeciej cechuje dominacja przemysłu lekkiego w gałęziowym profilu uprzemysłowienia. Duże znaczenie w tej klasie ma także przemysł elektromaszynowy.

Klasa czwarta obejmuje województwa, w których prawie równorzędną rolę odgrywają przemysły lekki i elektromaszynowy. Dużą wagę ma w tej klasie również przemysł mineralny i paliwowo-energetyczny.

Klasę piątą tworzą województwa, w których przoduje w gałęziowym profilu uprzemysłowienia przemysł spożywczy. Nieco mniejsze znaczenie mają w tej klasie przemysły elektromaszynowy i lekki. Niepoślednią rolę odgrywa także przemysł drzewno-papierniczy.

W kolejnych dwóch klasach pomimo, że w gałęziowym profilu uprzemysłowienia przeważa przemysł elektromaszynowy, to jednak występują między nimi zasadnicze różnice. Wynikają one z roli jaką w tych klasach odgrywają inne gałęzie przemysłu. W klasie szóstej wyróżnia się oprócz tego przemysł metalurgiczny i chemiczny, natomiast w klasie siódmej chemiczny i spożywczy.

W województwie chełmskim tworzącym klasę ósmą w gałęziowym profilu uprzemysłowienia góruje przemysł lekki i mineralny. Nikłe znaczenie mają przemysły: paliwowo-energetyczny, chemiczny i pozostałe gałęzie przemysłu. Nie występuje w tej klasie gałąź przemysłu metalurgicznego.

Klasę dziewiątą tworzy województwo gorzowskie, w którym w gałęziowym profilu uprzemysłowienia dominuje przemysł elektromaszynowy. Doniosłe znaczenie mają przemysły: chemiczny, drzewno-

-papierniczy, lekki i spożywczy, które odgrywają prawie równorzędną rolę. W klasie tej nie jest reprezentowana gałąź przemysłu metalurgicznego.

W kolejnej klasie jednoelementowej utworzonej przez województwo katowickie wybija się w gałęziowym profilu uprzemysłowienia przemysł paliwowo-energetyczny. Znaczny wpływ ma w tym województwie również przemysł elektromaszynowy i metalurgiczny, natomiast niewielkie znaczenie mają inne gałęzie przemysłu.

Województwo konińskie tworzące klasę jedenastą cechuje dominacja w gałęziowym profilu uprzemysłowienia przemysłu paliwowo-energetycznego. Duże znaczenie mają też w tej klasie następujące przemysły: elektromaszynowy, lekki i spożywczy.

W województwie legnickim na plan pierwszy wybija się przemysł metalurgiczny. Kolejno można wyróżnić w gałęziowym profilu uprzemysłowienia przemysły elektromaszynowy i lekki. Inne przemysły w tej klasie nie odgrywają poważniejszej roli.

Klasę trzynastą utworzoną przez województwo ostrołęckie cechuje przeważanie przemysłu elektromaszynowego i drzewno-papierniczego. Niemalże znaczenie ma również w tej klasie przemysł spożywczy.

Mianem klasy wyspecjalizowanej w danej gałęzi przemysłu będzie nazywana taka, której udział w całości wynosi co najmniej 25% .

W przemyśle paliwowo-energetycznym wyspecjalizowała się klasa dziesiąta i jedenasta, w przemyśle metalurgicznym - klasa dwunasta, w przemyśle elektromaszynowym - klasa pierwsza, druga, szósta, siódma i trzynasta, w przemyśle mineralnym - klasa ósma, w przemyśle drzewno-papierniczym - klasa trzynasta, w przemyśle lekkim zaś - klasa trzecia, czwarta i ósma. Nie wyspecjalizowała się żadna z klas w przemyśle chemicznym, spożywczym i pozostałych gałęziach przemysłu.

6.2.3. Rejestracja zmian typologicznych w strukturze gałęziowej zatrudnienia w przemyśle województw Polski.

Do pomiaru zmian jakie zaszły w dwóch klasyfikacjach województw Polski ze względu na podobieństwo skali w strukturze gałęziowej zatrudnienia z lat 1975 i 1980 zastosowano miarę zaproponowaną przez W. M. Randa. Miara ta policzona według wzoru /5.21/ wynosi $R = 0,8945$. Świadczy to o dużym podobieństwie przeprowadzonych klasyfikacji z lat 1975 i 1980. Można zatem wnioskować, że ustalone klasy charakteryzują się dużą stabilnością w czasie.

Składy klas: drugiej - 12. elementowej, piątej - 3. elementowej oraz wszystkich klas jednoelementowych /od siódmej do jedenastej/ pozostały niezmienione w podziale wynikowym w roku 1980. Niewielka zmiana nastąpiła w składzie klasy trzeciej od której w 1980 r. odłączyło się tylko województwo białostockie.

Zasadnicza zmiana nastąpiła w składzie klasy pierwszej, a klasy czwarta i szósta uległy likwidacji w roku 1980. Klasa pierwsza uległa rozbiciu, tworząc w roku 1980 trzony dwóch odrębnych klas: pierwszej /10 województw/ i trzeciej /7 województw/. Województwa klasy szóstej z podziału wynikowego z roku 1975 zostały przyłączone w 1980 r. przez klasę pierwszą, natomiast dwa województwa z klasy czwartej, tj. białkopodlaskie i łomżyńskie przyłączyły się do klasy trzeciej. Stan liczebny klasy pierwszej uzupełniło ponadto województwo białostockie. Województwo chełmskie w 1980 roku utworzyło klasę jednoelementową.

Jak z tego widać główne zmiany nastąpiły w składach klas zawierających województwa o bardzo niskim i niskim poziomie uprzemysłowienia.

Oczywiście - jak już wspomniano - na otrzymane rezultaty ^{naależy} spojrzeć z pewnym dystansem, jako że przeprowadzono badanie jednocechowe, które nie oddaje w pełni istoty problemu. Zbyt duża jest zdaniem autora różnica w poziomie uprzemysłowienia między województwem katowickim a następnymi w kolejności klasami, i to zarówno w

roku 1975 jak i 1980. Ponadto zakwalifikowanie niektórych województw do klas o niskim poziomie uprzemysłowienia może budzić poważne wątpliwości /np. województwo konińskie i płockie/. Wynikają one głównie z niedoskonałości badań poziomu uprzemysłowienia za pomocą miernika jakim jest zatrudnienie. Mimo swej prostoty posiada on wiele wad, które omówiono w podrozdziale 6.1. Zaniża się przede wszystkim poziom uprzemysłowienia tych województw, które mają dobrze rozwinięty przemysł paliwowo-energetyczny i /lub/ metalurgiczny /np. wcześniej wspomniane dwa województwa oraz województwo jeleniogórskie i wałbrzyskie/.

Miara Randa zastosowana do pomiaru zmian jakie zaszły w dwóch klasyfikacjach województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej zatrudnienia z lat 1975 i 1980 dała wynik $R = 0,7950$. Na podstawie wartości tej miary można wnioskować, że w 1980 r. w stosunku do roku 1975 zaszły dość istotne zmiany w klasyfikacjach województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej zatrudnienia.

Wprawdzie i w tym przypadku dużo klas charakteryzuje się stabilnością w czasie, ale są to klasy o małej liczebności. Składy dwóch klas 2. elementowych /klasa czwarta i szósta/ oraz pięciu 1. elementowych /klasa siódma oraz klasy od dziewiątej do dwunastej/ nie uległy zmianie w podziale wynikowym w roku 1980.

Zasadnicze zmiany nastąpiły w składach klas o dużej liczebności. Z klasy pierwszej wyodrębniła się grupa 10 województw tworząc w 1980 r. z województwem pilskim odrębną klasę, w której w gałęziowym profilu uprzemysłowienia zdecydowanie góruje przemysł elektromaszynowy. Pozostałe 8 województw klasy pierwszej wraz z 7. województwami klasy drugiej, 2. z klasy trzeciej /białostockie, zielonogórskie/ i województwo częstochowskie /tworzące w roku 1975 klasę jednoelementową/ utworzyło w 1980 r. nową klasę, w której przeważa przemysł elektromaszynowy, a duże znaczenie ma przemysł lekki i spożywczy.

Pozostałe województwa klasy drugiej, tj. przemyskie, suwalskie i zamojskie w roku 1980 utworzyły odrębną klasę. W wyniku rozpadu klasy trzeciej powstały w 1980 r. dwie odrębne klasy /trzecia - 4. elementowa i czwarta - 3. elementowa/, a pozostałe dwa województwa zostały przyłączone do klasy pierwszej. Województwo ostrołęckie w 1980 r. utworzyło klasę jednoelementową.

Cechą charakterystyczną obu klasyfikacji jest to, że znaczna liczba klas specjalizuje się w przemyśle elektromaszynowym /odpowiednio sześć i pięć klas/ i lekkim /odpowiednio dwie i trzy klasy/. W obu klasyfikacjach brak jest klas o specjalizacji w przemyśle chemicznym i pozostałych gałęziach przemysłu. Ponadto w roku 1975 nie wyspecjalizowała się żadna z klas w przemyśle drzewno-papierniczym, a w roku 1980 zaś - w przemyśle spożywczym.

6.3. Klasyfikacja województw Polski ze względu na strukturę gałęziową podstawowych relacji ekonomicznych w przemyśle.

6.3.1. Podobieństwo wielkości /skali/.

Rezultaty klasyfikacji hierarchicznej województw Polski ze względu na podobieństwo skali w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1975 ilustruje tabela 6.12. Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok e / $e = 1, 2, \dots, 47$ / jako pierwszy spełnia nierówność:

$$L_{e+1} > 0,4003 .$$

Proces klasyfikacji przerywa się więc po 36. iteracji, otrzymując ostatecznie podział zbioru województw na następujące klasy /w kolejności według liczebności/:

I. województwa: bydgoskie, gdańskie, kieleckie, krośnieńskie, lubelskie, opolskie, poznańskie, radomskie, rzeszowskie, szczeciń-

skie, tarnobrzeskie, tarnowskie, toruńskie, wrocławskie, wrocławskie;

II. województwa: białostockie, ciechanowskie, koszalińskie, leszczyńskie, nowosądeckie, olsztyńskie, ostrołęckie, pilskie, przemyskie, siedleckie, słupeckie, suwalskie, zamojskie;

III. województwa: elbląskie, kaliskie, piotrkowskie, skierniewickie, zielonogórskie;

IV. województwa: stołeczne warszawskie, bielskie;

V. województwa: białkopodlaskie, łomżyńskie;

VI. województwa: chełmskie, sieradzkie;

VII. województwa: jeleniogórskie, wałbrzyskie;

VIII. województwa: konińskie, płockie;

IX. województwa: miejskie krakowskie, legnickie;

X. województwo częstochowskie;

XI. województwo gorzowskie;

XII. województwo katowickie;

XIII. województwo miejskie łódzkie.

W wyodrębnionych klasach znajdują się województwa o zbliżonym gałęziowym poziomie uprzemysłowienia.

Graficzne wyniki klasyfikacji województw Polski ze względu na podobieństwo skali w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1975 prezentuje rys. 6.5. Ponadto - w celu ułatwienia interpretacji otrzymanych wyników - dla każdej gałęzi w wydzielonej klasie wyznaczono średnie arytmetyczne ze znormalizowanych wartości podstawowych relacji ekonomicznych w przemyśle i zaprezentowano je w tabeli 6.13. Wyznaczono również uporządkowanie liniowe województw według ogólnego poziomu uprzemysłowienia /por. tab. 6.14/, które posłuży do analizy wydzielonych klas.

Tabela 6.12

Rezultaty klasyfikacji województw Polski ze względu na podobieństwo skali
w strukturze gałęziowej podstawowych relacji ekonomicznych
w przemyśle /1975 r./

Nr iteracji	Nry połączonych klas	Odległość międzyklasowa	Nr iteracji	Nry połączonych klas	Odległość międzyklasowa
1	32 - 35	0, 1131	25	28 - 33	0, 2888
2	3 - 39	0, 1279	26	42 - 43	0, 2917
3	10 - 47	0, 1290	27	5 - 19	0, 3000
4	22 - 32	0, 1635	28	9 - 30	0, 3015
5	7 - 36	0, 1718	29	3 - 21	0, 3175
6	10 - 22	0, 1847	30	16 - 31	0, 3265
7	5 - 27	0, 1875	31	25 - 28	0, 3374
8	17 - 26	0, 1911	32	18 - 20	0, 3400
9	19 - 34	0, 1953	33	5 - 42	0, 3405
10	30 - 38	0, 2020	34	3 - 25	0, 3434
11	44 - 46	0, 2092	35	1 - 4	0, 3595
12	40 - 48	0, 2100	36	5 - 10	0, 3713
13	13 - 49	0, 2203	37	2 - 3	0, 4046
14	5 - 41	0, 2223	38	5 - 9	0, 4116
15	3 - 17	0, 2282	39	8 - 18	0, 4174
16	10 - 15	0, 2418	40	5 - 11	0, 4562
17	7 - 40	0, 2512	41	12 - 16	0, 4956
18	9 - 13	0, 2513	42	2 - 6	0, 5046
19	3 - 29	0, 2521	43	1 - 5	0, 5295
20	5 - 44	0, 2599	44	1 - 2	0, 5607
21	12 - 45	0, 2617	45	1 - 12	0, 6131
22	3 - 7	0, 2756	46	8 - 14	0, 6451
23	6 - 37	0, 2788	47	1 - 8	0, 6984
24	2 - 23	0, 2884	48	1 - 24	0, 7961

Źródło: obliczenia własne.

Rys. 6.5: Klasyfikacja województw Polski ze względu na podobieństwo skali w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1975.

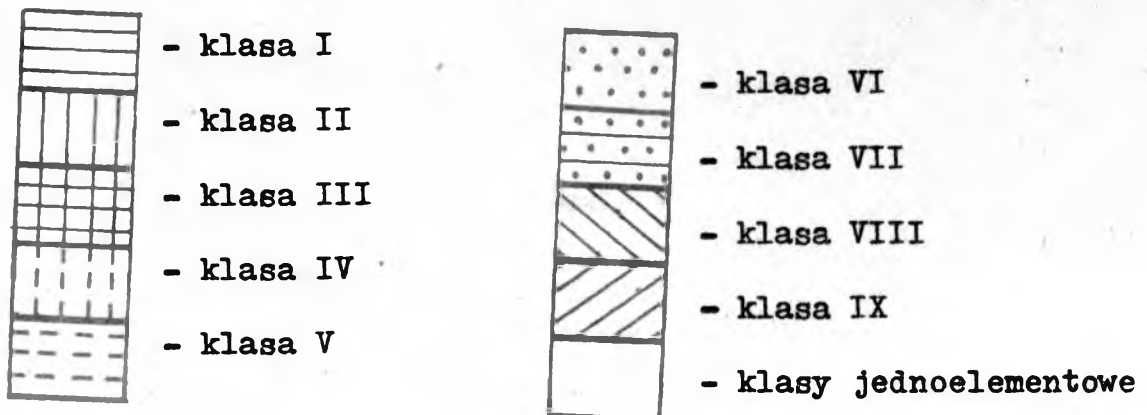


Tabela 6.13

Średnie arytmetyczne ze znormalizowanych /wg formuły /2.46// wartości podstawowych relacji ekonomicznych dla każdej gałęzi w wydzielonej klasie /1975 r./

Nr klasy \ Nr gałęzi	1	2	3	4	5	6	7	8	9
1	0,0411	0,0147	0,1772	0,0763	0,0358	0,0273	0,0347	0,0855	0,0150
2	0,0220	0,0000	0,0627	0,0090	0,0227	0,0319	0,0345	0,0823	0,0140
3	0,0165	0,0014	0,1244	0,0258	0,0283	0,0266	0,1024	0,0734	0,0167
4	0,0511	0,0518	0,4100	0,0899	0,0307	0,0319	0,1263	0,0702	0,0463
5	0,0057	0,0000	0,0293	0,0007	0,0077	0,0124	0,0529	0,0523	0,0090
6	0,0099	0,0000	0,0308	0,0034	0,1037	0,0104	0,0936	0,0632	0,0046
7	0,1729	0,0018	0,1412	0,0426	0,0953	0,0538	0,2091	0,0536	0,0104
8	0,3011	0,0278	0,0938	0,0168	0,0084	0,0071	0,0276	0,0970	0,0154
9	0,0514	0,3653	0,1373	0,0382	0,0513	0,0193	0,0513	0,0871	0,0178
10	0,0348	0,1409	0,1231	0,0231	0,0799	0,0471	0,1090	0,0363	0,0189
11	0,0209	0,0000	0,0771	0,1149	0,0166	0,0767	0,0527	0,0642	0,0153
12	0,6396	0,3032	0,3116	0,0781	0,0911	0,0188	0,0447	0,0650	0,0085
13	0,0787	0,0005	0,2838	0,1375	0,0340	0,0428	0,8429	0,1061	0,0307

Źródło: obliczenia własne.

Wartości miary poziomu uprzemysłowienia /1975 r./

Lp.	Województwo	Wartości miary
1	katowickie	1,5610
2	miejskie łódzkie	1,5583
3	stołeczne warszawskie	0,9126
4	bielskie	0,9047
5	miejskie krakowskie	0,8527
6	jeleniogórskie	0,7937
7	legnickie	0,7864
8	wałbrzyskie	0,7683
9	płockie	0,6756
10	wrocławskie	0,6342
11	częstochofskie	0,6133
12	opolskie	0,6128
13	tarnowskie	0,5498
14	bydgoskie	0,5443
15	tarnobrzefskie	0,5340
16	szczecińskie	0,5312
17	gdańskie	0,5172
18	konińskie	0,5152
19	kieleckie	0,5148
20	toruńskie	0,5018
21	poznańskie	0,4898
22	krośnieńskie	0,4843
23	zielonogórskie	0,4772
24	kaliskie	0,4671
25	lubelskie	0,4451
26	piotrkowskie	0,4398
27	gorzowskie	0,4388
28	radomskie	0,4385
29	rzeszowskie	0,4132
30	włocławskie	0,4076
31	chełmskie	0,3652
32	elbląskie	0,3476
33	skierniewickie	0,3474
34	łofczyńskie	0,3356
35	olsztyńskie	0,3242
36	ślupskie	0,3230
37	pilskie	0,3192
38	białostockie	0,3125
39	koszalińskie	0,2938
40	przemyskie	0,2904
41	ostrolęckie	0,2795
42	sieradzkie	0,2745
43	suwalskie	0,2564
44	nowosądeckie	0,2500
45	zamojskie	0,2346
46	siedleckie	0,2116
47	ciechanowskie	0,2010
48	łomżyńskie	0,1753
49	białskopodlaskie	0,1656

Poziom uprzemysłowienia:

bardzo niski i niski	ponizej 0,3602
średni	(0,3602 ; 0,5039)
znaczny	(0,5039 ; 0,6476)
wysoki i bardzo wysoki	powyzej 0,6476

Źródło: obliczenia własne.

Najwyższy poziom uprzemysłowienia przedstawiają klasy dwunasta i trzynasta, przy czym województwo katowickie przeważa pod względem poziomu uprzemysłowienia - wśród wydzielonych klas - w przemyśle paliwowo-energetycznym, natomiast województwo miejskie łódzkie w przemyśle lekkim, chemicznym i spożywczym. O poziomie uprzemysłowienia klasy dwunastej decyduje w głównej mierze przemysł paliwowo-energetyczny oraz przemysły elektromaszynowy i metalurgiczny, klasy trzynastej zaś - przemysł lekki i elektromaszynowy.

Trzy klasy dwuelementowe, a mianowicie czwarta, siódma i dziewiąta prezentują wysoki poziom uprzemysłowienia. W klasie czwartej największy wpływ na poziom uprzemysłowienia ma przemysł elektromaszynowy, w klasie siódmej - przemysł lekki, paliwowo-energetyczny i elektromaszynowy, w klasie dziewiątej zaś - przemysł metalurgiczny i elektromaszynowy. Klasa czwarta - wśród wydzielonych klas - przeważa pod względem poziomu uprzemysłowienia w przemyśle elektromaszynowym, natomiast klasa dziewiąta - w przemyśle metalurgicznym.

Bardzo niski poziom uprzemysłowienia prezentuje klasa piąta, w której spośród gałęzi wyróżnia się zdecydowanie przemysł lekki i spożywczy. Nie występuje w tej klasie gałąź przemysłu metalurgicznego.

Klasa druga przedstawia niski poziom uprzemysłowienia. O poziomie uprzemysłowienia tej klasy decyduje przemysł spożywczy i elektromaszynowy. Nie jest reprezentowana w niej gałąź przemysłu metalurgicznego.

Województwa chełmskie i sieradzkie tworzące klasę szóstą prezentują odpowiednio średni i niski poziom uprzemysłowienia. Na ogólny poziom uprzemysłowienia województw tej klasy największy wpływ mają przemysły mineralny i lekki. Klasa szоста przoduje - wśród wydzielonych klas - w przemyśle mineralnym. Nie występuje w tej klasie przemysł metalurgiczny.

Klasa trzecia utworzona jest z województw o niskim i średnim poziomie uprzemysłowienia. Największy wpływ na ogólny poziom

uprzemysłowienia tej klasy mają przemysły elektromaszynowy i lekki.

Średni poziom uprzemysłowienia przedstawia klasa jedenasta, o którym w znacznej mierze decydują przemysły: chemiczny, drzewno-papierniczy i elektromaszynowy. Klasa ta ponadto - wśród wydzielonych klas - przeważa pod względem poziomu uprzemysłowienia w przemyśle drzewno-papierniczym.

Klasa pierwsza utworzona jest z województw o średnim i znacznym poziomie uprzemysłowienia. Decydujący wpływ w tej klasie mają przemysły: elektromaszynowy, spożywczy i chemiczny.

Województwo częstochowskie tworzące klasę jednoelementową o numerze dziesięć prezentuje znaczny poziom uprzemysłowienia, o którym decydują przemysły: metalurgiczny, elektromaszynowy i lekki.

Województwa płockie i konińskie przedstawiające odpowiednio wysoki i znaczny poziom uprzemysłowienia utworzyły klasę ósmą, w której spośród innych gałęzi wyróżnia się zdecydowanie przemysł paliwowo-energetyczny.

Województwa znajdujące się w jednej klasie klasyfikacji na ogół zajmują zbliżone pozycje po uszeregowaniu ich według wartości poziomu uprzemysłowienia.

Rezultaty klasyfikacji hierarchicznej województw Polski ze względu na podobieństwo skali w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1980 prezentuje tabela 6.15. Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok $e / e = 1, 2, \dots, 47/$ jako pierwszy spełnia nierówność:

$$L_{e+1} > 0,3954 .$$

Proces klasyfikacji przerywa się więc po 37. kroku, otrzymując w efekcie podział zbioru województw na następujące klasy:

I. województwa: białskopodlaskie, białostockie, ciechanowskie, elbląskie, koszalińskie, leszczyńskie, nowosądeckie, olsztyńskie,

ostrołęckie, pilskie, przemyskie, siedleckie, słupekie, suwalskie, zamojskie;

II. województwa: bydgoskie, gorzowskie, krośnieńskie, opolskie, radomskie, szczecińskie, tarnobrzeckie, tarnowskie, toruńskie, włocławskie;

III. województwa: bielskie, gdańskie, kieleckie, lubelskie, poznańskie, rzeszowskie, wrocławskie;

IV. województwa: częstochowskie, kaliskie, piotrkowskie, skierniewickie, zielonogórskie;

V. województwa: chełmskie, sieradzkie;

VI. województwa: jeleniogórskie, wałbrzyskie;

VII. województwa: konińskie, plockie;

VIII. województwa: miejskie krakowskie, legnickie;

IX. województwo stołeczne warszawskie;

X. województwo katowickie;

XI. województwo łomżyńskie;

XII. województwo miejskie łódzkie.

W wydzielonych klasach znajdują się województwa o zbliżonym gałęziowym poziomie uprzemysłowienia.

Graficzne rezultaty klasyfikacji województw Polski ze względu na podobieństwo skali w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1980 ilustruje rys. 6.6. Ponadto - w celu ułatwienia interpretacji otrzymanych wyników - dla każdej gałęzi w wydzielonej klasie wyznaczono średnie arytmetyczne ze znormalizowanych wartości podstawowych relacji ekonomicznych w przemyśle i zaprezentowano je w tabeli 6.16. Wyznaczono również uporządkowanie liniowe województw według ogólnego poziomu uprzemysłowienia /por. tab. 6.17/, które posłuży do analizy wydzielonych klas.

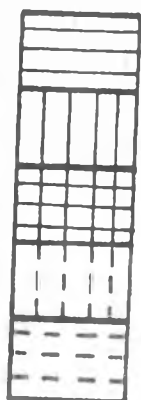
Tabela 6.15

Rezultaty klasyfikacji województw Polski ze względu na podobieństwo skali
w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle /1980 r./

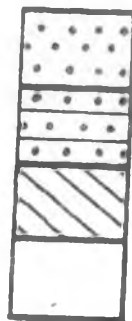
Nr iteracji	Nry połączonych klas	Odległość międzyklasowa	Nr iteracji	Nry połączonych klas	Odległość międzyklasowa
1	32 - 35	0,1132	25	5 - 34	0,2703
2	3 - 39	0,1168	26	3 - 33	0,2851
3	10 - 32	0,1383	27	6 - 37	0,2968
4	10 - 47	0,1549	28	2 - 40	0,3032
5	7 - 36	0,1581	29	3 - 21	0,3101
6	5 - 44	0,1765	30	5 - 43	0,3289
7	3 - 9	0,1838	31	2 - 3	0,3384
8	10 - 22	0,1964	32	4 - 10	0,3449
9	17 - 26	0,2048	33	5 - 11	0,3566
10	13 - 49	0,2084	34	2 - 28	0,3655
11	40 - 48	0,2093	35	18 - 20	0,3691
12	19 - 42	0,2234	36	8 - 13	0,3824
13	3 - 29	0,2260	37	16 - 31	0,3829
14	7 - 17	0,2289	38	4 - 5	0,4040
15	34 - 41	0,2388	39	4 - 8	0,4094
16	13 - 30	0,2396	40	2 - 23	0,4516
17	3 - 25	0,2410	41	2 - 6	0,4568
18	5 - 27	0,2426	42	4 - 12	0,4677
19	10 - 15	0,2457	43	2 - 4	0,5403
20	5 - 19	0,2559	44	1 - 14	0,5792
21	11 - 46	0,2592	45	2 - 16	0,5881
22	3 - 7	0,2623	46	1 - 18	0,6179
23	12 - 45	0,2648	47	1 - 2	0,7189
24	13 - 38	0,2673	48	1 - 24	0,7732

źródło: obliczenia własne.

Rys. 6.6. Klasyfikacja województw Polski ze względu na podobieństwo skali w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1980.



- klasa I
- klasa II
- klasa III
- klasa IV
- klasa V



- klasa VI
- klasa VII
- klasa VIII
- klasy jednoelementowe

Tabela 6.16

Średnie arytmetyczne ze znormalizowanych /wg formuły /2.16// wartości podstawowych relacji ekonomicznych dla każdej gałęzi w wydziolonej klasie /1980 r./

Nr klasy	Nr gałęzi	1	2	3	4	5	6	7	8	9
1		0,0155	0,0000	0,0987	0,0089	0,0215	0,0342	0,0379	0,0757	0,0210
2		0,0487	0,0126	0,1516	0,1035	0,0399	0,0364	0,0412	0,0846	0,0178
3		0,0294	0,0225	0,2798	0,0477	0,0316	0,0256	0,0592	0,0831	0,0192
4		0,0177	0,0256	0,1475	0,0298	0,0331	0,0364	0,1175	0,0702	0,0227
5		0,0083	0,0000	0,0466	0,0037	0,0855	0,0112	0,0946	0,0731	0,0076
6		0,1462	0,0016	0,1725	0,0413	0,0986	0,0491	0,2111	0,0504	0,0146
7		0,2678	0,0211	0,1228	0,0219	0,0106	0,0084	0,0466	0,0874	0,0264
8		0,0484	0,3728	0,1686	0,0383	0,0457	0,0177	0,0567	0,0848	0,0276
9		0,0536	0,0460	0,5667	0,0684	0,0463	0,0232	0,0448	0,0765	0,0487
10		0,6757	0,3827	0,3633	0,0722	0,0856	0,0166	0,0432	0,0621	0,0122
11		0,0054	0,0000	0,0308	0,0024	0,0055	0,0396	0,0735	0,0583	0,0129
12		0,0752	0,0006	0,3166	0,1225	0,0290	0,0328	0,7960	0,0946	0,0362

Źródło: obliczenia własne.

Wartości miary poziomu uprzemysłowienia /1980 r./

Lp.	Województwo	Wartości miary
1	katowickie	1,7139
2	miejskie łódzkie	1,5039
3	stołeczne warszawskie	0,9745
4	bielskie	0,8978
5	legnickie	0,8969
6	miejskie krakowskie	0,8248
7	wałbrzyskie	0,7882
8	jeleniogórskie	0,7836
9	płockie	0,7327
10	wrocławskie	0,6578
11	opolskie	0,6567
12	częstochowskie	0,6114
13	tarnobrzeskie	0,5955
14	bydgoskie	0,5924
15	gdańskie	0,5794
16	szczecińskie	0,5638
17	kieleckie	0,5582
18	tarnowskie	0,5344
19	toruńskie	0,5330
20	krośnieńskie	0,5281
21	poznańskie	0,5169
22	zielonogórskie	0,5151
23	rzeszowskie	0,5119
24	piotrkowskie	0,5000
25	kaliskie	0,4971
26	konińskie	0,4942
27	gorzowskie	0,4863
28	radomskie	0,4839
29	lubelskie	0,4675
30	włocławskie	0,3920
31	skierniowickie	0,3812
32	śląskie	0,3790
33	łoszczyńskie	0,3759
34	elbląskie	0,3759
35	białostockie	0,3634
36	chełmskie	0,3600
37	pilskie	0,3592
38	koszalińskie	0,3565
39	olsztyńskie	0,3150
40	sieradzkie	0,3022
41	przemyskie	0,3012
42	ostrołęckie	0,2943
43	nowosądeckie	0,2785
44	suwalskie	0,2720
45	ciechanowskie	0,2601
46	siedleckie	0,2449
47	łomżyńskie	0,2288
48	zamojskie	0,2226
49	białskopodlaskie	0,2114

Poziom uprzemysłowienia :

bardzo niski i niski	poniżej 0,3908
średni	(0,3908 0,5362)
znaczny	(0,5362 0,6816)
wysoki i bardzo wysoki	powyżej 0,6816

Źródło: obliczenia własne.

Najwyższy poziom uprzemysłowienia prezentuje województwo katowickie i miejskie łódzkie, przy czym - wśród wydzielonych klas - pod względem poziomu uprzemysłowienia przeważa pierwsze z nich w przemyśle paliwowo-energetycznym i metalurgicznym, drugie zaś - w przemyśle lekkim, chemicznym i spożywczym. O poziomie uprzemysłowienia klasy dziesiątej decyduje przede wszystkim przemysł paliwowo-energetyczny, metalurgiczny i elektromaszynowy, klasy dwunastej zaś - przemysł lekki i elektromaszynowy.

Dwie klasy dwuelementowe, tj. szósta i ósma oraz jednoelementowa klasa dziewiąta prezentują wysoki poziom uprzemysłowienia. W klasie szóstej największy wpływ na ogólny poziom uprzemysłowienia ma przemysł lekki, elektromaszynowy i paliwowo-energetyczny, w klasie ósmej - przemysł metalurgiczny i elektromaszynowy, w klasie dziewiątej zaś - przemysł elektromaszynowy. Klasa szósta - wśród wydzielonych klas - przeważa pod względem poziomu uprzemysłowienia w przemyśle mineralnym i drzewno-papierniczym, natomiast klasa dziewiąta - w przemyśle elektromaszynowym.

Klasy: pierwsza, piąta i jedenasta skupiają województwa o bardzo niskim i niskim poziomie uprzemysłowienia, na który decydujący wpływ mają przemysły: odpowiednio elektromaszynowy i spożywczy; lekki, mineralny i spożywczy oraz lekki i spożywczy. Nie jest reprezentowana w tych klasach gałąź przemysłu metalurgicznego.

Klasy: druga, trzecia i czwarta skupiają województwa o średnim i znacznym poziomie uprzemysłowienia. Jedynie w klasie trzeciej województwo bielskie prezentuje wysoki poziom uprzemysłowienia, a w klasie czwartej województwo skierniewickie - niski. Różnice między tymi klasami wynikają z roli jaką odgrywają w nich poszczególne gałęzie przemysłu. W klasie drugiej o ogólnym poziomie uprzemysłowienia decyduje przemysł elektromaszynowy i chemiczny, w klasie trzeciej - przemysł elektromaszynowy, w klasie czwartej zaś - przemysł elektromaszynowy i lekki.

Województwa płockie i konińskie przedstawiające odpowiednio

wysoki i średni poziom uprzemysłowienia utworzyły klasę siódmą, w której spośród innych gałęzi wyróżnia się zdecydowanie przemysł paliwowo-energetyczny.

Województwa znajdujące się w jednej klasie klasyfikacji na ogół zajmują zbliżone pozycje po uszeregowaniu ich według wartości poziomu uprzemysłowienia.

6.3.2. Podobieństwo kształtu /formy/.

Wyniki klasyfikacji hierarchicznej województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1975 prezentuje tabela 6.18. Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok $e / e = 1, 2, \dots, 47/$ jako pierwszy spełnia nierówność:

$$L_{e+1} > 0,2926 .$$

Proces klasyfikacji przerywa się więc po 35. iteracji, otrzymując w efekcie podział zbioru obiektów na następujące klasy:

I. województwa: białkopodlaskie, białostockie, bydgoskie, elbląskie, kaliskie, nowosądeckie, opolskie, pilskie, śląskie, szczecińskie, zielonogórskie;

II. województwa: stołeczne warszawskie, gdańskie, kieleckie, lubelskie, poznańskie, rzeszowskie, wrocławskie;

III. województwa: ciechanowskie, koszalińskie, leszczyńskie, olsztyńskie, siedleckie, suwalskie, zamojskie;

IV. województwa: gorzowskie, tarnobrzeskie, tarnowskie, toruńskie, włocławskie;

V. województwa: bielskie, piotrkowskie, skierniewickie;

VI. województwa: chełmskie, sieradzkie;

VII. województwa: jeleniogórskie, wałbrzyskie;

Tabela 6.18

Rezultaty klasyfikacji województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle / 1975 r. /

Nr iteracji	Nry połączonych klas	Odległość międzyklasowa	Nr iteracji	Nry połączonych klas	Odległość międzyklasowa
1	10 - 35	0,0901	25	6 - 37	0,2187
2	1 - 47	0,1015	26	5 - 25	0,2231
3	10 - 32	0,1046	27	12 - 45	0,2262
4	7 - 21	0,1057	28	11 - 44	0,2272
5	3 - 39	0,1092	29	4 - 30	0,2364
6	9 - 13	0,1202	30	5 - 9	0,2367
7	1 - 10	0,1241	31	42 - 43	0,2541
8	7 - 36	0,1460	32	28 - 33	0,2616
9	40 - 48	0,1485	33	2 - 5	0,2630
10	2 - 3	0,1519	34	11 - 42	0,2741
11	44 - 46	0,1523	35	23 - 24	0,2896
12	1 - 22	0,1595	36	2 - 7	0,2965
13	25 - 27	0,1642	37	14 - 16	0,2985
14	18 - 20	0,1723	38	12 - 19	0,3062
15	30 - 38	0,1779	39	8 - 18	0,3214
16	2 - 49	0,1857	40	2 - 28	0,3253
17	19 - 34	0,1873	41	4 - 23	0,3389
18	7 - 17	0,1874	42	1 - 2	0,3664
19	5 - 29	0,1922	43	12 - 14	0,3849
20	7 - 40	0,1983	44	1 - 11	0,4095
21	9 - 41	0,2048	45	4 - 6	0,4281
22	1 - 15	0,2086	46	1 - 12	0,4598
23	16 - 31	0,2126	47	1 - 4	0,4711
24	7 - 26	0,2148	48	1 - 8	0,5634

Źródło: obliczenia własne.

Rys. 6.7. Klasyfikacja województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1975.

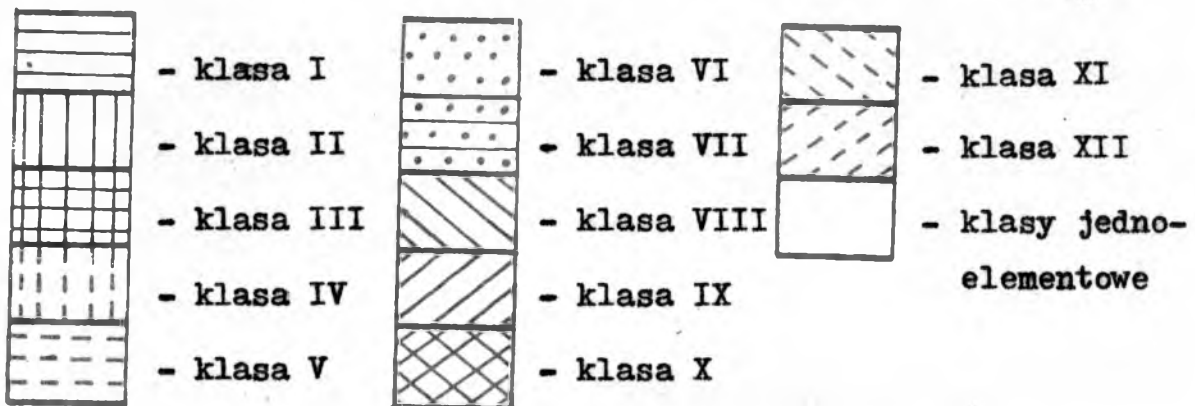


Tabela 6.19

Średnie arytmetyczne ze znormalizowanych /wg formuły /2,45// wartości podstawowych relacji ekonomicznych dla każdej gałęzi w wydzielonej klasie /1975 r./

Nr klasy	Nr gałęzi	Nr klasy								
		1	2	3	4	5	6	7	8	9
1	a	0,2959	0,0258	0,9246	0,1949	0,2636	0,3070	0,4983	0,6853	0,1264
	b	8,91	0,78	27,83	5,87	7,93	9,24	15,00	20,63	3,81
2	a	0,1351	0,0712	1,0000	0,1818	0,1466	0,0866	0,1177	0,3372	0,0350
	b	6,34	3,34	46,92	8,53	6,88	4,07	5,52	15,82	2,58
3	a	0,1853	0,0000	0,5833	0,1135	0,2171	0,2957	0,2911	0,9574	0,1214
	b	6,70	0	21,10	4,11	7,85	10,69	10,53	34,63	4,39
4	a	0,1766	0,0930	0,7074	0,8198	0,1619	0,2483	0,2277	0,5261	0,1154
	b	5,74	3,02	23,00	26,65	5,27	8,07	7,40	17,10	3,75
5	a	0,1554	0,0524	0,8627	0,4454	0,2178	0,1374	0,8384	0,3453	0,1082
	b	4,91	1,66	27,27	14,08	6,89	4,34	26,51	10,92	3,42
6	a	0,0833	0,0000	0,2449	0,0280	0,7868	0,0808	0,7166	0,4658	0,0370
	b	3,41	0	10,02	1,15	32,20	3,31	29,33	19,07	1,51
7	a	0,6428	0,0081	0,5598	0,1551	0,3702	0,2036	0,8016	0,2163	0,0401
	b	21,44	0,27	18,68	5,17	12,35	6,79	26,74	7,22	1,34
8	a	0,9041	0,1253	0,4116	0,0744	0,0442	0,0396	0,1665	0,3880	0,0605
	b	40,83	5,66	18,59	3,36	2,00	1,79	7,52	12,52	2,73
9	a	0,1366	0,9575	0,4636	0,1261	0,1645	0,0637	0,1831	0,2597	0,0638
	b	5,65	39,59	19,17	5,21	6,80	2,63	7,57	10,74	2,64
10	a	0,5993	0,0006	0,8246	0,3479	0,2712	0,1252	0,2550	0,3651	0,0382
	b	21,20	0,02	29,17	12,31	9,59	4,43	9,02	12,91	1,35
11	a	0,0974	0,0003	0,3130	0,0959	0,0700	0,0882	1,0000	0,4457	0,0731
	b	4,46	0,01	14,33	4,39	3,21	4,04	45,80	20,41	3,35
12	a	0,5805	0,0006	0,5802	0,0267	0,3307	0,4994	0,3193	0,7251	0,1889
	b	17,85	0,02	17,85	0,82	10,17	15,36	9,82	22,30	5,81
13	a	0,2116	0,8651	0,7617	0,1437	0,4819	0,2908	0,6746	0,2263	0,1178
	b	5,61	22,92	20,18	3,81	12,77	7,71	17,88	6,00	3,12
14	a	1,0000	0,5326	0,5253	0,1350	0,1409	0,0309	0,0789	0,1203	0,0151
	b	38,78	20,65	20,37	5,23	5,46	1,20	3,06	4,66	0,59

a - w liczbach bezwzględnych,

b - w %.

Źródło: obliczenia własne.

- VIII. województwa: konińskie, płockie;
- IX. województwa: miejskie krakowskie, legnickie;
- X. województwa: krośnieńskie, radomskie;
- XI. województwa: łomżyńskie, miejskie łódzkie;
- XII. województwa: ostrołęckie, przemyskie;
- XIII. województwo częstochowskie;
- XIV. województwo katowickie.

W wyodrębnionych klasach znajdują się województwa o zbliżonym gałęziowym profilu uprzemysłowienia.

Graficzne rezultaty klasyfikacji województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1975 prezentuje rys. 6.7. Ponadto - w celu ułatwienia interpretacji otrzymanych wyników - dla każdej gałęzi w wydzielonej klasie wyznaczono średnie arytmetyczne ze znormalizowanych wartości podstawowych relacji ekonomicznych w przemyśle i zaprezentowano je w tabeli 6.19.

W gałęziowym profilu uprzemysłowienia województw tworzących klasę pierwszą uwidacznia się przemysł elektromaszynowy. Niepoślednie znaczenie ma w tej klasie również przemysł spożywczy i lekki.

Charakterystyczną cechą klas: drugiej, ósmej, dziewiątej, jedenastej i czternastej jest to, że obejmują województwa w których zdecydowanie góruje w gałęziowym profilu uprzemysłowienia jedna gałąź przemysłu w stosunku do pozostałych. W klasie drugiej gałęzią tą jest przemysł elektromaszynowy, w klasie ósmej i czternastej - przemysł paliwowo-energetyczny, w klasie dziewiątej - przemysł metalurgiczny, a w klasie jedenastej - przemysł lekki. Dużą rolę odgrywa w klasie drugiej przemysł spożywczy, w klasie ósmej i jedenastej - przemysł elektromaszynowy i spożywczy, w klasie dziewiątej - przemysł elektromaszynowy, a w klasie czternastej - przemysł metalurgiczny i elektromaszynowy.

Województwa klasy trzeciej cechuje dominacja przemysłu spożywczego w gałęziowym profilu uprzemysłowienia. Duże znaczenie w tej klasie ma także przemysł elektromaszynowy. W województwach tej klasy nie występuje przemysł metalurgiczny.

Klasę czwartą tworzą województwa, w których przoduje w gałęziowym profilu uprzemysłowienia przemysł chemiczny oraz elektromaszynowy. Niepoślednią rolę odgrywa również przemysł spożywczy.

Klasa piąta obejmuje województwa, w których równorzędną rolę odgrywają przemysły elektromaszynowy i lekki. Dużą wagę ma w tej klasie także przemysł chemiczny.

W klasie szóstej w gałęziowym profilu uprzemysłowienia góruje przemysł mineralny i lekki, natomiast nikłe znaczenie mają przemysły: paliwowo-energetyczny, chemiczny, drzewno-papierniczy i pozostałe gałęzie przemysłu. Nie występuje w tej klasie gałąź przemysłu metalurgicznego.

Województwa jeleniogórskie i wałbrzyskie tworzące klasę siódmą cechuje dominacja w gałęziowym profilu uprzemysłowienia przemysłu lekkiego. Nieco mniejszą rolę odgrywa w tej klasie przemysł paliwowo-energetyczny i elektromaszynowy.

W klasie dziesiątej na plan pierwszy wybijają się przemysły elektromaszynowy. Kolejno wyróżniają się w gałęziowym profilu uprzemysłowienia przemysły: paliwowo-energetyczny, chemiczny i spożywczy.

Klasa dwunasta obejmuje województwa, w których przeważa w gałęziowym profilu uprzemysłowienia przemysł spożywczy. Nieco mniejszą rolę odgrywa przemysł paliwowo-energetyczny i elektromaszynowy.

Przemysłami odgrywającymi dużą rolę w klasie trzynastej są: metalurgiczny, elektromaszynowy i lekki.

Mianem klasy wyspecjalizowanej w danej gałęzi przemysłu będzie nazywana taka, której udział w całości wynosi co najmniej 25% .

W przemyśle paliwowo-energetycznym wyspecjalizowała się klasa ósma i czternasta, w przemyśle metalurgicznym - klasa dziewiąta, w przemyśle elektromaszynowym - klasa pierwsza, druga, piąta i

dziesiąta, w przemyśle chemicznym - klasa czwarta, w przemyśle mineralnym - klasa szósta, w przemyśle lekkim - klasa piąta, szósta, siódma i jedenasta, w przemyśle spożywczym zaś - klasa trzecia. Nie wyspecjalizowała się żadna z klas w przemyśle drzewno-papierniczym i w pozostałych gałęziach przemysłu.

Rezultaty klasyfikacji hierarchicznej województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1980 przedstawia tabela 6.20. Z ciągu klasyfikacji wybiera się tę, dla której odpowiadający jej krok e / $e = 1, 2, \dots, 47$ / jako pierwszy spełnia nierówność:

$$L_{e+1} > 0,2909 .$$

Proces klasyfikacji przerywa się zatem po 35. kroku, otrzymując w wyniku podział zbioru województw na następujące klasy:

I. województwa: białkopodlaskie, białostockie, bielskie, elbląskie, kaliskie, nowosądeckie, piotrkowskie, skierniewickie, słupskie, wałbrzyskie, zielonogórskie;

II. województwa: ciechanowskie, koszalińskie, leszczyńskie, olsztyńskie, przemyskie, siedleckie, suwalskie, szczecińskie, zamojskie;

III. województwa: stołeczne warszawskie, gdańskie, kieleckie, lubelskie, pilskie, poznańskie, rzeszowskie, wrocławskie;

IV. województwa: bydgoskie, gorzowskie, krośnieńskie, opolskie, tarnobrzeskie, tarnowskie, toruńskie, włocławskie;

V. województwa: chełmskie, sieradzkie;

VI. województwa: miejskie krakowskie, legnickie;

VII. województwa: płockie, radomskie;

VIII. województwo częstochowskie;

IX. województwo jeleniogórskie;

Tabela 6.20

Rezultaty klasyfikacji województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle / 1980 r./

Nr iteracji	Nry połączonych klas	Odległość międzyklasowa	Nr iteracji	Nry połączonych klas	Odległość międzyklasowa
1	10 - 32	0,0850	25	5 - 11	0,2233
2	2 - 39	0,0883	26	18 - 20	0,2321
3	35 - 47	0,0954	27	7 - 40	0,2322
4	7 - 21	0,0993	28	31 - 34	0,2405
5	1 - 35	0,1054	29	5 - 43	0,2413
6	2 - 13	0,1125	30	4 - 45	0,2450
7	3 - 49	0,1207	31	7 - 33	0,2530
8	1 - 10	0,1241	32	2 - 38	0,2540
9	2 - 9	0,1321	33	6 - 37	0,2556
10	1 - 22	0,1492	34	2 - 4	0,2622
11	7 - 36	0,1539	35	5 - 19	0,2838
12	17 - 26	0,1597	36	14 - 16	0,2965
13	40 - 48	0,1630	37	8 - 18	0,3016
14	44 - 46	0,1643	38	1 - 2	0,3049
15	2 - 25	0,1734	39	1 - 7	0,3300
16	7 - 17	0,1777	40	12 - 24	0,3423
17	19 - 42	0,1790	41	14 - 31	0,3538
18	5 - 44	0,1830	42	1 - 5	0,3707
19	2 - 3	0,1833	43	6 - 23	0,3764
20	4 - 30	0,1838	44	1 - 28	0,3979
21	1 - 15	0,2052	45	6 - 12	0,4104
22	7 - 41	0,2059	46	1 - 14	0,4734
23	19 - 27	0,2071	47	1 - 6	0,4913
24	1 - 29	0,2166	48	1 - 8	0,5347

Źródło: obliczenia własne.

Rys. 6.8. Klasyfikacja województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1980.

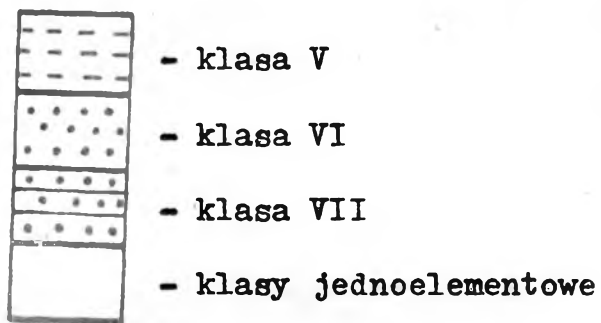
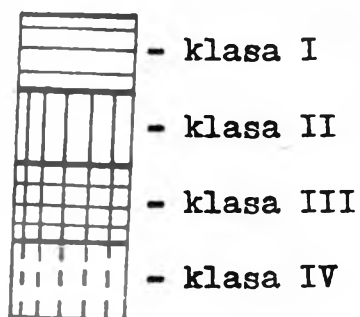


Tabela 6.21

Średnie arytmetyczne ze znormalizowanych /wg formuły /2.45// wartości podstawowych relacji ekonomicznych dla każdej gałęzi w wydzielonej klasie /1980 r./

Nr klasy	Nr gałęzi	Nr gałęzi								
		1	2	3	4	5	6	7	8	9
1	a	0,2441	0,0178	0,9703	0,1557	0,2111	0,2397	0,6305	0,5310	0,1364
	b	7,78	0,57	30,94	4,96	6,73	7,64	20,10	16,93	4,35
2	a	0,2197	0,0046	0,7014	0,1335	0,2091	0,2542	0,2851	0,8722	0,1818
	b	7,68	0,16	24,51	4,67	7,31	8,88	9,96	30,48	6,35
3	a	0,1404	0,0605	1,0000	0,1456	0,1573	0,1112	0,1131	0,3201	0,0853
	b	6,58	2,84	46,87	6,83	7,37	5,21	5,30	15,00	4,00
4	a	0,2783	0,0748	0,7966	0,7039	0,2614	0,2522	0,2340	0,4538	0,1011
	b	8,82	2,37	25,24	22,30	8,28	7,99	7,42	14,38	3,20
5	a	0,0969	0,0000	0,3834	0,0304	0,7647	0,0921	0,7327	0,5975	0,0631
	b	3,51	0	13,89	1,10	27,70	3,34	26,54	21,64	2,28
6	a	0,1506	0,9394	0,5113	0,1277	0,1452	0,0495	0,1679	0,2423	0,0842
	b	6,23	38,85	21,14	5,28	6,01	2,05	6,94	10,02	3,48
7	a	0,6707	0,0011	0,7043	0,1676	0,1101	0,0416	0,1771	0,3157	0,0472
	b	30,00	0,05	31,51	7,50	4,93	1,86	7,92	14,12	2,11
8	a	0,1414	0,7304	0,8943	0,1444	0,2019	0,2808	0,6095	0,2171	0,1607
	b	4,18	21,61	26,46	4,27	5,97	8,31	18,03	6,42	4,75
9	a	0,6519	0,0008	0,4324	0,2225	0,3462	0,1934	0,7758	0,1349	0,0571
	b	23,16	0,03	15,36	7,90	12,30	6,87	27,56	4,79	2,03
10	a	0,9885	0,5972	0,5705	0,1151	0,1262	0,0251	0,0636	0,1021	0,0175
	b	37,91	22,90	21,88	4,41	4,84	0,96	2,52	3,91	0,67
11	a	1,0000	0,2491	0,3963	0,1053	0,0770	0,0434	0,4038	0,4859	0,2216
	b	33,53	8,35	13,29	3,53	2,58	1,46	13,54	16,29	7,43
12	a	0,1160	0,0000	0,3649	0,0290	0,0743	0,5188	0,9501	0,7202	0,1441
	b	3,98	0	12,51	0,99	2,54	17,78	32,57	24,69	4,94
13	a	0,1300	0,0007	0,3910	0,1592	0,0392	0,0385	1,0000	0,1177	0,0448
	b	6,77	0,04	20,35	8,29	2,04	2,00	52,05	6,13	2,33
14	a	0,5422	0,0019	0,6936	0,0256	0,3490	0,8665	0,0682	0,6476	0,1437
	b	16,24	0,06	20,78	0,77	10,45	25,96	2,04	19,40	4,30

a - w liczbach bezwzględnych,
b - w %.

Źródło: obliczenia własne.

- X. województwo katowickie;
- XI. województwo konińskie;
- XII. województwo łomżyńskie;
- XIII. województwo miejskie łódzkie;
- XIV. województwo ostrołęckie.

W wyodrębnionych klasach znajdują się województwa o zbliżonym gałęziowym profilu uprzemysłowienia.

Graficzne rezultaty klasyfikacji województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle w roku 1980 prezentuje rys. 6.8. Ponadto - w celu ułatwienia interpretacji otrzymanych wyników - dla każdej gałęzi w wydzielonej klasie wyznaczono średnie arytmetyczne ze znormalizowanych wartości podatawowych relacji ekonomicznych w przemyśle i zaprezentowano je w tabeli 6.21.

W gałęziowym profilu uprzemysłowienia województw tworzących klasę pierwszą uwidacznia się przemysł elektromaszynowy. Niepoślednie znaczenie ma w tej klasie również przemysł lekki.

Województwa klasy drugiej cechuje dominacja przemysłu spożywczego w gałęziowym profilu uprzemysłowienia. Duże znaczenie ma w tej klasie także przemysł elektromaszynowy. W klasie tej nie jest reprezentowana gałąź przemysłu metalurgicznego.

Znamienną cechą klas: trzeciej, szóstej, dziesiątej, jedenastej i trzynastej jest to, że skupiają one województwa, w których zdecydowanie góruje w profilu uprzemysłowienia jedna gałąź przemysłu w stosunku do pozostałych. W klasie trzeciej gałęzią tą jest przemysł elektromaszynowy, w klasie szóstej - przemysł metalurgiczny, w klasie dziesiątej i jedenastej - przemysł paliwowo-energetyczny oraz w klasie trzynastej - przemysł lekki. Dużą rolę odgrywa również w klasie trzeciej i jedenastej przemysł spożywczy, w klasie szóstej i trzynastej - przemysł elektromaszynowy, a w klasie dziesiątej - przemysł metalurgiczny i elektromaszynowy.

Klasę czwartą tworzą województwa, w których przoduje w gałęziowym profilu uprzemysłowienia przemysł elektromaszynowy oraz chemiczny.

W klasie piątej w gałęziowym profilu uprzemysłowienia góruje przemysł mineralny i lekki, natomiast nikłe znaczenie mają przemysły: paliwowo-energetyczny, chemiczny, drzewno-papierniczy i pozostałe gałęzie przemysłu. Nie występuje w tej klasie gałąź przemysłu metalurgicznego.

Klasa siódma obejmuje województwa, w których przeważają przemysły elektromaszynowy i paliwowo-energetyczny. Dużą wagę ma w tej klasie także przemysł spożywczy.

W województwie częstochowskim tworzącym klasę ósmą dominują przemysły: elektromaszynowy, metalurgiczny i lekki.

W gałęziowym profilu uprzemysłowienia województwa jeleniogórskiego przeważa przemysł lekki oraz paliwowo-energetyczny. Duże znaczenie ma też w tym województwie przemysł elektromaszynowy.

Województwo łomżyńskie tworzące klasę dwunastą cechuje dominacja w gałęziowym profilu uprzemysłowienia przemysłu lekkiego. Nieco mniejszą rolę odgrywa w tej klasie przemysł spożywczy i drzewno-papierniczy.

W klasie czternastej przeważa w gałęziowym profilu uprzemysłowienia przemysł drzewno-papierniczy. Nieco mniejsze znaczenie w tej klasie mają przemysły: elektromaszynowy, spożywczy i paliwowo-energetyczny.

Mianem klasy wyspecjalizowanej w danej gałęzi przemysłu będzie nazywana taka, której udział w całości wynosi co najmniej 25% .

W przemyśle paliwowo-energetycznym wyspecjalizowała się klasa siódma i jedenasta, w przemyśle metalurgicznym - klasa szósta, w przemyśle elektromaszynowym - klasa pierwsza, trzecia, czwarta, siódma i ósma, w przemyśle mineralnym - klasa piąta, w przemyśle drzewno-papierniczym - klasa czternasta, w przemyśle lekkim - klasa piąta, dziewiąta, dwunasta i trzynasta, w przemyśle spożyw-

czym zaś - klasa druga. Nie wyspecjalizowała się żadna z klas w przemyśle chemicznym i w pozostałych gałęziach przemysłu.

6.3.3. Rejestracja zmian typologicznych w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle województw Polski.

Do pomiaru zmian jakie zaszły w dwóch klasyfikacjach województw Polski ze względu na podobieństwo skali w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle z lat 1975 i 1980 zastosowano miarę W. M. Randa o postaci /5.21/, otrzymując wartość $R = 0,9098$. Świadczy to o dużym podobieństwie przeprowadzonych klasyfikacji z lat 1975 i 1980. Można zatem wnioskować, że ustalone klasy charakteryzują się dużą stabilnością w czasie.

Składy klas: czterech dwuelementowych /klasy od szóstej do dziewiątej/ oraz dwóch jednoelementowych /klasa dwunasta i trzynasta/ nie uległy zmianie w podziale wynikowym w roku 1980.

Niewielka zmiana nastąpiła w składzie klasy drugiej, do której w roku 1980 przyłączyły się województwa białkopodlaskie i gorzowskie oraz w składzie klasy trzeciej, z której na miejsce województwa elbląskiego przyłączyło się województwo częstochowskie, tworzące w roku 1975 klasę jednoelementową.

Zasadnicza zmiana nastąpiła w składzie klasy pierwszej 15. elementowej. Klasa ta uległa rozbiciu, tworząc w roku 1980 trzony dwóch odrębnych klas: drugiej /9 województw/ i trzeciej /6 województw/. Skład klasy drugiej uzupełniło województwo gorzowskie tworzące w roku 1975 klasę jednoelementową, a do składu klasy trzeciej przybyło w roku 1980 województwo bielskie.

Po odłączeniu się z klasy czwartej województwa bielskiego oraz z klasy piątej województwa białkopodlaskiego w roku 1980 powstały dwie nowe klasy jednoelementowe utworzone przez pozostałe w tych klasach województwa.

Miara Randa zastosowana do pomiaru zmian jakie zaszły w dwóch klasyfikacjach województw Polski ze względu na podobieństwo kształtu w strukturze gałęziowej podstawowych relacji ekonomicznych w przemyśle z lat 1975 i 1980 dała wynik $R = 0,9073$. Świadczy to o dużym podobieństwie przeprowadzonych klasyfikacji z lat 1975 i 1980. Można zatem wnioskować, że ustalone klasy charakteryzują się dużą stabilnością w czasie.

Nie uległy zmianie w 1980 r. składy tylko czterech klas, tzn. dwóch dwuelementowych /piątej i szóstej/ oraz dwóch jednoelementowych /ósmej i dziesiątej/. W obu klasyfikacjach nie zaszły jednak istotne zmiany, ponieważ składy klas o dużej liczebności charakteryzowały się stabilnością w czasie.

Z klasy pierwszej w 1980 r. odłączyły się cztery województwa i tyleż się ich przyłączyło na skutek przyłączenia klasy piątej /województwa: bielskie, piotrkowskie i skierniewickie/ oraz przyłączenia województwa wałbrzyskiego.

Do składu klas: drugiej, trzeciej i czwartej dołączyły w 1980 r. odpowiednio: województwo pilskie; województwo przemyskie i szczecińskie oraz województwo bydgoskie, krośnieńskie i opolskie.

Wskutek rozpadu pięciu klas dwuelementowych /siódmej, ósmej, dziesiątej, jedenastej i dwunastej/ powstało w 1980 r. pięć nowych klas jednoelementowych /utworzonych przez województwa: jeleńskie, konińskie, łomżyńskie, miejskie łódzkie, ostrołęckie/ oraz nowa klasa dwuelementowa /województwa płockie i radomskie/.

Cechą charakterystyczną obu klasyfikacji jest to, że znaczna liczba klas specjalizuje się w przemyśle elektromaszynowym /odpowiednio 4 i 5 klas/ i lekkim /4 klasy w obydwu podziałach/.

W obu klasyfikacjach brak jest klas o specjalizacji w pozostałych gałęziach przemysłu. Ponadto w roku 1975 nie wyspecjalizowała się żadna z klas w przemyśle drzewno-papierniczym, a w roku 1980 zaś - w przemyśle chemicznym.

7. WNIOSKI KOŃCOWE

Przeprowadzone badania dotyczące klasyfikacji województw Polski ze względu na podobieństwo gałęziowego poziomu oraz profilu uprzemysłowienia potwierdziły przydatność zastosowanych metod statystyczno-matematycznych. Uzyskano spodziewane rezultaty, a wyniki tych badań odpowiadały logiczno-intuicyjnej interpretacji obserwowanych zjawisk.

W związku z powyższym można pokusić się o sformułowanie pewnych ogólnych wniosków natury metodycznej oraz wynikających z badań empirycznych.

7.1. Wnioski natury metodycznej.

Wnioski te można z kolei podzielić na dwie grupy:

1. praktyczne korzyści wynikające z zastosowania do tego typu badań metod klasyfikacji,
2. wnioski wynikające z nie rozwiązania w aspekcie teoretycznym pewnych problemów związanych z metodami klasyfikacji.

Jeżeli chodzi o korzyści jakie wynikają z zastosowania metod klasyfikacji do tego typu badań to można do nich zaliczyć następujące:

a/ sposoby prezentacji wielkości /skali/ oraz kształtu /formy/ struktur /por. rozdział drugi/ umożliwiają badanie podobieństwa struktur obiektów ze względu na poziom oraz profil wybranego zjawiska;

b/ miary podobieństwa struktur /por. rozdział trzeci/ pozwalają określić stopień podobieństwa badanych obiektów jednocześnie ze względu na szeroki zestaw cech ustalonych dla wielu elementów struktury. W efekcie można przeprowadzić wielocechową klasyfikację badanych obiektów;

c/ w jednostkach o podobnej strukturze badanego zjawiska łatwiej jest dokonać ekonometrycznej analizy zjawisk ekonomicznych. Zatem przeprowadzenie klasyfikacji powinno stanowić czynność wstępną, poprzedzającą modelowanie ekonometryczne;

d/ wyodrębnienie klas obiektów "jednorodnych" pozwala nam na uchwycenie elementów charakterystycznych, które odróżniają je od innych klas. Analiza bez uprzedniej klasyfikacji, w przypadku dużej liczby obiektów badania, jest praktycznie niemożliwa, a nawet jeśli badacz jest znawcą przedmiotu badania, jest bardzo pracochłonna. Z pomocą elektronicznej techniki obliczeniowej przy wykorzystaniu algorytmów metod klasyfikacji można szybko otrzymać pożądany efekt badania;

e/ wyodrębnienie klas obiektów o podobnej strukturze badanego zjawiska stanowić może cenne narzędzie i informację przy podejmowaniu decyzji w sprawie stanu i perspektyw rozwoju obiektów gospodarczych /województw/;

f/ za pomocą miar podobieństwa podziałów wynikowych /por. rozdział piąty/ można dokonać rejestracji zmian w czasie w klasyfikacjach obiektów ze względu na obrazy ich struktur.

Do wniosków wynikających z nie rozwiązania w sensie teoretycznym pewnych problemów związanych z metodami klasyfikacji można zaliczyć następujące:

a/ występuje problem doboru odpowiedniego zestawu cech diagnostycznych, które stosowane będą w danym badaniu. Jest on o tyle ważny, że przyjęte cechy w znacznym stopniu decydują o ostatecznych rezultatach. Istnieje wprawdzie wiele podejść do rozwiązania tego

zagadnienia oraz bardzo dużo technik statystycznych, jednakże nie rozwiązują one w sposób zadowalający tego problemu;

b/ na pytanie "która z miar podobieństwa powinna być użyta w badaniu" nie znajdujemy dotychczas jednoznacznej odpowiedzi. Istnieje bardzo wiele miar podobieństwa, natomiast brak jest kryteriów formalnych, w oparciu o które można by zdecydować się na wybór jednej z nich. Autor w rozdziale trzecim znacznie ogranicza liczbę miar podobieństwa możliwych do zastosowania w badaniach strukturalnych poprzez wprowadzenie dodatkowego warunku /por. rozdział 3.1 warunek 5^o/. Wnioski wysunięte przez autora na końcu tego rozdziału również zmniejszają liczbę miar podobieństwa. Nie rozwiązuje to jednak istniejącego problemu wyboru miary podobieństwa struktur. Powstaje oczywiście pytanie czy problem ten jest możliwy do rozwiązania w ogóle i czy zachodzi konieczność jego rozwiązania;

c/ w związku z tym, że jest bardzo dużo metod klasyfikacji /wybrane metody zostały omówione w rozdziale czwartym/ należy zdecydować się na wybór jednej z nich, tj. takiej, która byłaby najlepsza według określonych kryteriów. Na ogół wykorzystuje się w tym celu kryteria merytoryczne, analizując wady i zalety tychże metod. Takie postępowanie nie pozwala na jednoznaczny wybór, ale zawęża nieco liczbę metod klasyfikacji, które powinno stosować się w badaniu. Wykorzystuje się również w tym celu pewne kryteria formalne /patrz podrozdział 4.1/. Wydaje się, że łączne rozpatrywanie kryteriów merytorycznych i formalnych pozwoli na dokonanie wyboru wąskiej grupy metod optymalnych dla danego problemu badawczego;

d/ kolejny problem, wiążący się z metodami klasyfikacji, polega na ustaleniu optymalnej liczby klas na jakie należy podzielić zbiór badanych obiektów. Pojawiają się tutaj intuicyjnie dwa pytania: co to jest dobry podział wynikowy oraz czy podział wynikowy P_h jest lepszy od podziału P_h . Na pytania te obecnie nie można udzielić jednoznacznej odpowiedzi. Przegląd sposobów wyznaczania optymalnej liczby klas przedstawiono w podrozdziale 4.3. W związku

z powyższym od badacza problemu wymaga się dwojakiego rodzaju umiejętności, tzn. w zakresie wybranej dyscypliny badawczej /merytoryczna znajomość zagadnienia/ oraz w zakresie opanowania metod statystyczno-ekonometrycznych /metodologiczna znajomość zagadnienia/, wyrażające naukową swobodę badacza;

e/ jeśli chodzi o problem podobieństwa dwóch podziałów wynikowych, to autor uważa, że warunki przez niego postawione w rozdziale piątym /o ile zostaną zaakceptowane/ rozwiązują go w decydującym stopniu.

7.2. Wnioski wynikające z badań empirycznych.

Przeprowadzone badania empiryczne /por. rozdział szósty/ potwierdziły celowość badań struktur gałęziowych przemysłu przy pomocy metod klasyfikacji. Wyniki badań empirycznych pozwalają na sformułowanie kilku wniosków natury ogólnej:

1. Województwa Polski są silnie zróżnicowane ze względu na gałęziowy poziom oraz profil uprzemysłowienia i to zarówno w roku 1975 jak i 1980.
2. W okresie 1975 - 1980 nie nastąpiły zasadnicze zmiany w klasyfikacjach województw Polski ze względu na gałęziowy profil oraz poziom uprzemysłowienia /wniosek ten wyciągnięty został na podstawie badań wielocechowych, albowiem badania jednocechowe mają ograniczony zakres jak wykazano w podrozdziale 6.1/. Przeprowadzone badania pozwoliły określić zmiany struktury gałęziowej przemysłu województw mianem umiarkowanych. Przemawia za tym duża stabilność w czasie wyodrębnionych klas obiektów podobnych zarówno ze względu na gałęziowy poziom jak i profil uprzemysłowienia.
3. Porównując /przy pomocy miary Randa/ wyniki klasyfikacji wielocechowej województw Polski ze względu na gałęziowy poziom oraz profil uprzemysłowienia z wynikami klasyfikacji jednocechowej

można wysnuć wniosek, że występują między nimi dość istotne różnice i to zarówno przy badaniu poziomu jak i profilu uprzemysłowienia dla lat 1975 i 1980 /miara Randa w czterech analizowanych sytuacjach przyjmowała wartości od 0,80 do 0,84/.

4. Otrzymane rezultaty przesądziły o celowości wydzielenia klas województw podobnych ze względu na gałęziowy profil oraz poziom uprzemysłowienia. Na podstawie niniejszej pracy można sformułować propozycję dalszych studiów w tym zakresie z uwzględnieniem dłuższego okresu badania /przeobrażenia strukturalne charakteryzują się powolnością zmian/. Gdy lista okresów badania będzie dostatecznie długa można pokusić się o przeprowadzenie analizy dynamicznej, w celu wyodrębnienia podokresów, dla których podziały wynikowe nieistotnie będą różnić się od siebie.

L I T E R A T U R A

1. R. L. Ackoff, Decyzje optymalne w badaniach stosowanych, Warszawa 1969, PWN.
2. S. A. Ajvazjan, Z. I. Beżaeva, O. V. Staroverov, Klassifikacija mnogomiernych nabludenij, Moskwa 1974, Statistika.
3. J. Akerman, L'aspect structurel, "Revue Economique", 1954, nr 6.
4. M. R. Anderberg, Cluster analysis for applications, New York, San Francisco, London 1973, Academic Press.
5. W. R. Atchley, E. H. Bryant, Multivariate statistical methods: among-groups covariation, Stroudsburg, Pennsylvania 1975, Halsted Press.
6. S. Bartosiewicz, Ekonometria. Technologia ekonometrycznego przetwarzania informacji, Warszawa 1976, PWE.
7. E. J. Bijnen, Cluster analysis, The Netherlands 1973, Tilburg University Press.
8. T. Borys, Elementy teorii jakości, Warszawa 1980, PWN.
9. T. Borys, Kategoria jakości w statystycznej analizie porównawczej, Prace Naukowe AE we Wrocławiu, nr 284, 1984.
10. W. Borysiuk, Próba kwantyfikacji stopnia zmian struktury produkcji polskiego przemysłu w okresie 1950 - 1970, "Gospodarka Planowa", 1973, nr 5.
11. W. Bukietyński, Z. Hellwig, U. Królik, A. Smoluk, Uwagi o dyskryminacji zbiorów skończonych, Prace Naukowe AE we Wrocławiu, 1969, nr 21.
12. Z. Chojnicki, T. Czyż, Metody taksonomii numerycznej w regionalizacji geograficznej, Warszawa 1973, PWN.
13. S. Chomątowski, A. Sokołowski, Taksonomia struktur, "Przegląd

- Statystyczny", 1978, nr 2.
14. S. Chomątowski, K. Kukuła, Metoda oceny podobieństwa oraz zmian struktury gospodarczej /na przykładzie województw Polski w latach 1963 - 1973/, Zeszyty Naukowe AE w Krakowie, 1980, nr 127.
 15. M. Cieślak, Modele zapotrzebowania na kadry kwalifikowane, Warszawa 1976, PWN.
 16. M. Cieślak, Dobór syndromu zmiennych do porządkowania liniowego obiektów wielowymiarowych, Referat wygłoszony na konferencji naukowej nt. "Taksonomia struktur" , Zachełmie 1984.
 17. R. M. Cormack, A review of classification /with discussion/, Journal of the Royal Statistical Society, 1971, Series A, Volume 134, Part 3.
 18. H. Cramer, metody matematyczne w statystyce, Warszawa 1958, PWN.
 19. M. Dąbrowski, K. Laus-Mączyńska, Metody wyszukiwania i klasyfikacji informacji, Warszawa 1978, WNT.
 20. B. S. Duran, P. L. Odell, Cluster analysis. A survey, Berlin, Heidelberg, New York 1974, Springer-Verlag.
 21. B. S. Everitt, Cluster analysis, London 1977, Heinemann Educational Books Ltd.
 22. B. S. Everitt, The analysis of contingency tables, London 1979, Chapman and Hall.
 23. J. S. Farris, On the cophenetic correlation coefficient, "Systematic Zoology", 1969, 18/3/, s. 279-285.
 24. J. Fierich, Próba zastosowania metod taksonomicznych do rejonizacji systemów rolniczych w województwie krakowskim, "Myśl Gospodarcza", 1957, nr 1.
 25. L. Fisher, J. W. Van Ness, Admissible clustering procedures, "Biometrika", 1971, 58.1.
 26. F. Flamant, Structure économique et périodes longues, "Revue Economique", 1954, nr 6.
 27. K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus, S. Zubrzy-

- oki, Taksonomia Wrocławska, "Przegląd Antropologiczny",
Poznań 1951, t.17.
28. P. S. Florence, Brytyjski i amerykański system przemysłowy,
Warszawa 1965, PWN.
29. J. J. Fortier, H. Solomon, Clustering procedures, w: P. R. Krishnaiah /ed./, Multivariate analysis, New York, London 1966,
Academic Press.
30. M. M. Goodman, Distance analysis in biology, "Systematic
Zoology", 1972, 21/2/.
31. L. A. Goodman, W. H. Kruskal, Measures of association for cross
classifications, New York, Heidelberg 1979, Springer-Verlag.
32. G. Gorzelak, Zastosowanie statystycznej analizy porównawczej
do badania przestrzennej struktury rolnictwa w Polsce /według
województw/, Praca doktorska, Wrocław 1980.
33. T. Grabiński, Metody analizy zbieżności wyników dyskryminacji
zbiorów, Zeszyty Naukowe AE w Krakowie, 1980, nr 127.
34. T. Grabiński, S. Wydymus, A. Zeliaś, Metody doboru zmiennych
w modelach ekonometrycznych, Warszawa 1982, PWN.
35. T. Grabiński, S. Wydymus, A. Zeliaś, Metody prognozowania
rozwoju społeczno-gospodarczego, Warszawa 1983, PWE.
36. J. P. Guilford, Podstawowe metody statystyczne w psychologii
i pedagogice, Warszawa 1964, PWN.
37. J. A. Hartigan, Clustering algorithms, New York, London,
Sydney, Toronto 1975, John Wiley & Sons.
38. Z. Hellwig, Elementy rachunku prawdopodobieństwa i statystyki
matematycznej, Warszawa 1975, PWN.
39. Z. Hellwig, Wielowymiarowa analiza porównawcza i jej zastosowanie
w badaniach wielocechowych obiektów gospodarczych, w:
W. Welfe /red./, Metody i modele ekonomiczno-matematyczne w
doskonaleniu zarządzania gospodarką socjalistyczną, Warszawa
1981, PWE.
40. Z. Hellwig, Zastosowanie metody taksonomicznej do typologiczne-

- go podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr, "Przegląd Statystyczny", 1968, z.4.
41. M. Hussain, Taksonomiczne metody podziału zbiorów skończonych, Praca doktorska, Kraków 1982.
 42. K. Jajuga, Metody analizy wielowymiarowej w ilościowych badaniach przestrzennych, Praca doktorska, Wrocław 1981.
 43. K. Jajuga, O sposobach określania ilości klas w zagadnieniach klasyfikacji i klasyfikacji rozmytej, Prace Naukowe AE we Wrocławiu, 1984, nr 262.
 44. N. Jardine, R. Sibson, Mathematical taxonomy, London, New York, Sydney, Toronto 1971, John Wiley & Sons Ltd.
 45. N. Jardine, R. Sibson, The construction of hierarchic and non-hierarchic classifications, "The Computer Journal", 1968, Volume 11, Number 2.
 46. A. Karpiński, Polityka zmian strukturalnych w gospodarce - jej cele i narzędzia, "Gospodarka Planowa" , 1974, nr 2.
 47. J. Kolonko, Analiza dyskryminacyjna i jej zastosowanie w ekonomii, Warszawa 1980, PWN.
 48. L. Król, Uwagi na marginesie teorii struktury, "Ekonomista", 1963, nr 3.
 49. P. Kubiak, Porównanie wybranych metod dyskryminacji zbiorów skończonych, Prace Naukowe AE we Wrocławiu, 1982, nr 203.
 50. J. Kucharczyk, Algorytmy analizy skupień w języku Algol 60, Warszawa 1982, PWN.
 51. K. Kukuła, Propozycja w zakresie pewnych miar dynamiki struktury, "Przegląd Statystyczny", 1975, nr 3.
 52. O. Kyn, P. Pelikan, Cybernetyka a ekonomia, Warszawa 1967, PWE.
 53. J. Lhomme, Matériaux pour une théorie de la structure économique, "Revue Economique", 1954, nr 6.
 54. J. Lisikiewicz, Wpływ postępu technicznego na zmiany strukturalne w przemyśle, Zeszyty Naukowe SGPiS, 1971, nr 83.

55. J. Lisikiewicz, J. Macieja, Rozwój strukturalny polskiego przemysłu, "Ekonomista", 1970, nr 3.
56. Mała Encyklopedia Ekonomiczna, Warszawa 1974, PWE.
57. A. Marchal, Systémes et structures économiques, Paris 1959, PUF.
58. S. Marciniak, Proporcje i struktura gospodarki socjalistycznej, Warszawa 1976, PWE.
59. S. Marciniak, Struktura produkcji a dynamika wzrostu gospodarczego, Warszawa 1970, PWN.
60. E. Marczewski, H. Steinhaus, O odległości systematycznej biotopów, "Zastosowania Matematyki", Warszawa - Wrocław 1958 - 1959, Tom IV, Zeszyt 9, PWN.
61. Metody ilościowe i modele w geografii, praca zbiorowa pod red. Z. Chojnickiego, Warszawa 1977, PWN.
62. Metody matematyczne i taksonomiczne w badaniach struktury przestrzennej rolnictwa, Praca zbiorowa, Warszawa 1970, Biuletyn KPZK PAN, Zeszyt 61.
63. B. Minc, Postęp ekonomiczny, Warszawa 1967, PWN.
64. B. Minc, Zarys systemu ekonomii politycznej, Warszawa 1970, PWN.
65. R. Mojena, Hierarchical grouping methods and stopping rules: an evaluation, "The Computer Journal", 1977, Volume 20, No. 4.
66. E. Nowak, Metodyka statystycznych analiz porównawczych efektywności obiektów rolniczych, Prace Naukowe AE we Wrocławiu, 1985, nr 292.
67. M. Opałko, Mierniki rozwoju regionów, Warszawa 1972, PWE.
68. W. Ostasiewicz, Dyskryminacja, klasyfikacja, rozpoznawanie, Prace Naukowe AE we Wrocławiu, 1980, nr 165.
69. J. Pamuła, A. Sokołowski, Propozycja wyznaczania podziału wynikowego aglomeracyjnych algorytmów taksonomicznych, Zeszyty Naukowe AE w Krakowie, 1980, nr 127.
70. J. J. Parysek, L. Wojtasiewicz, Metody analizy regionalnej i metody planowania regionalnego, Studia KPZK, Warszawa 1979, Tom LXIX, PWN.

71. J. Parysek, Struktura przestrzenna przemysłu regionu i jej zmiany, Poznań 1976, UAM.
72. J. Perkal, Taksonomia Wrocławska, "Przegląd Antropologiczny", Poznań 1953, T. 19.
73. F. Perroux, Comptes de la nation, Paris 1949.
74. J. Piaget, Strukturalizm, Warszawa 1972, Wiedza Powszechna.
75. W. Pluta, Grafowa metoda klasyfikacji cech II, Prace Naukowe WSE we Wrocławiu, 1972, z.35.
76. W. Pluta, Metoda oceny wyników delimitacji, Prace Naukowe AE we Wrocławiu, 1984, nr 262.
77. W. Pluta, Metody wielowymiarowej analizy porównawczej w modelowaniu informacji ekonomicznej kombinatu przemysłowego, Prace Naukowe AE we Wrocławiu, 1979, nr 156.
78. W. Pluta, Taksonomia izotoniczna i izomorficzna, Referat na II konferencję taksonomiczną, Wrocław 1981.
79. W. Pluta, Wielowymiarowa analiza porównawcza w badaniach ekonomicznych, Warszawa 1977, PWE.
80. J. Pocięcha, Kryteria oceny procedur taksonomicznych, "Przegląd Statystyczny", 1982, nr 1/2.
81. B. Podolec, K. Zając, Ekonometryczne metody ustalania rejonów konsumpcji, Warszawa 1978, PWE.
82. J. B. Racine, H. Raymond, Analiza ilościowa w geografii, Warszawa 1977, PWN.
83. W. M. Rand, Objective criteria for the evaluation of clustering methods, "JASA", Volume 66, Number 336.
84. C. R. Rao, Cluster analysis applied to a study of race mixture in human populations, w: J. Van Ryzin /ed./, Classification and clustering, New York, San Francisco, London 1977, Academic Press.
85. B. B. Rozin, Teoria rozpoznawania obrazów w badaniach ekonomicznych, Warszawa 1979, PWN.
86. S. Róg, Mierniki i metody badania stopnia uprzemysłowienia w

- układach regionalnych, w: Mierniki rozwoju regionów, Biblioteka Wiadomości Statystycznych, Warszawa 1969, Tom 9, GUS.
87. K. Secomski, Zmiany w rozmieszczeniu przemysłu polskiego na tle dynamiki jego wzrostu, "Nowe Drogi", 1963, nr 6/169/.
88. K. Sieklucki, Geometria z elementami topologii i algebry liniowej, Warszawa 1976, PWN.
89. P. H. A. Sneath, R. R. Sokal, Numerical taxonomy. The principles and practice of numerical classification, San Francisco 1973, W. H. Freeman.
90. R. R. Sokal, F. J. Rohlf, The comparison of dendrograms by objective methods, "Taxon", 1962, 11/2/.
91. J. Steczkowski, A. Zeliaś, Analiza wariacyjna i kowariancyjna w badaniach ekonomicznych, Warszawa 1982, PWN.
92. D. Strahl, Dyskryminacja zbioru według kryterium stabilnego podobieństwa, "Przegląd Statystyczny", 1982, nr 1/2.
93. D. Strahl, Modelowanie zjawisk złożonych. Modele infrastruktury społecznej, Prace Naukowe AE we Wrocławiu, 1980, nr 158.
94. D. Strahl, O prognozowaniu wektora cech, "Przegląd Statystyczny", 1978, nr 1.
95. M. Strihafka, Pojęcie struktury gospodarczej, Zeszyty Naukowe AE w Krakowie, 1981, nr 134.
96. Struktura branżowa i przestrzenna przemysłu 1975 i 1980, Warszawa 1982, GUS.
97. Struktura produkcji przemysłowej. Metody badania i kierunki zmian, Praca zbiorowa pod red. J. Lisikiewicza, Warszawa 1977, PWE.
98. F. A. Szczotka, Podstawy taksonomii numerycznej, Warszawa 1975, Problem węzłowy 11.2.1, Grupa tematyczna O3, Temat A.1, maszynopis, IGiPZ PAN.
99. C. Szmigiel, Modyfikacje wskaźnika zgodności kryteriów podziału, Prace Naukowe AE we Wrocławiu, 1981, nr 190.
100. C. Szmigiel, Wskaźnik zgodności kryteriów podziału, "Przegląd

- Statystyczny", 1976, nr 4.
101. Z. Szymła, Problemy rozwoju przestrzennych struktur przemysłowych, Zeszyty Naukowe AE w Krakowie, 1977, nr 40, Monografie.
 102. Z. Szymła, Klasyfikacja wielocechowa uprzemysłowienia województw, Zeszyty Naukowe WSE w Krakowie, 1972, nr 48.
 103. J. Tober, Niektóre metodologiczne problemy analiz strukturalnych w przemyśle, "Gospodarka Planowa", 1970, nr 7.
 104. J. W. Van Ness, Admissible clustering procedures, "Biometrika", 1973, 60.2.
 105. E. Wagemann, Introduction à la theorie du mouvement des affaires, Paris 1932.
 106. M. Walesiak, Kilka uwag o niektórych hierarchicznych metodach klasyfikacji, Referat wygłoszony na konferencji naukowej nt. "Taksonomia struktur", Zachełmie 1984.
 107. M. Walesiak, Metoda oceny podobieństwa struktur /na przykładzie struktury gałęziowej zatrudnienia w przemyśle uspołecznionym województw Polski w roku 1980/, "Wiadomości Statystyczne", 1982, nr 10.
 108. M. Walesiak, Miary odległości struktur udziałowych, Prace Naukowe AE we Wrocławiu, W druku.
 109. M. Walesiak, Podobieństwo wielkości /skali/ oraz kształtu /formy/ w złożonych badaniach strukturalnych, "Wiadomości Statystyczne", 1983, nr 3.
 110. M. Walesiak, Pojęcie, klasyfikacja i wskaźniki podobieństwa struktur gospodarczych, Prace Naukowe AE we Wrocławiu, 1984, nr 285.
 111. M. Walesiak, Propozycja rodziny miar odległości struktur udziałowych, "Wiadomości Statystyczne", 1983, nr 10.
 112. M. Walesiak, Propozycja wskaźnika odległości struktur prostych udziałowych, Prace Naukowe AE we Wrocławiu, 1985, nr 281.
 113. M. Walesiak, Sposoby rejestracji zmian w czasie w wynikach

- klasyfikacji, Referat wygłoszony na III Seminarium naukowym im. Z. Pawłowskiego, Karpacz 1985.
114. W. T. Williams, G. N. Lance, M. B. Dale, H. T. Clifford, Controversy concerning the criteria for taxonomic strategies, "The Computer Journal", 1971, Volume 14, Number 2.
 115. T. Wójcik, Zarys teorii klasyfikacji, Warszawa 1965, PWN.
 116. S. Wydymus, Metody taksonomii numerycznej w modelowaniu ekonometrycznym, Referat wygłoszony na I Seminarium naukowym im. Z. Pawłowskiego, Kozubnik 1983.
 117. S. Wydymus, Metody wielowymiarowej analizy rozwoju społeczno-gospodarczego, Zeszyty Naukowe AE w Krakowie, Seria Specjalna: Monografie, 1984, nr 62.
 118. S. Wydymus, Taksonometryczne mierniki tempa rozwoju społeczno-gospodarczego, Prace Naukowe AE we Wrocławiu, 1984, nr 262.
 119. L. Zacher, Zmiany strukturalno-jakościowe w gospodarce w dobie rewolucji naukowo-technicznej, Warszawa 1977, PWN.
 120. K. Zając, Zarys metod statystycznych, Warszawa 1982, PWE.
 121. S. Zawadzki, Analiza struktury przestrzennej przemysłu Polski Ludowej, Warszawa 1965, Studia KPZK, Tom X, PWN.
 122. J. Żebrok, Metody badania struktury gałęziowej przemysłu, Zeszyty Naukowe AE w Krakowie, 1975, nr 75.