RUCH PRAWNICZY, EKONOMICZNY I SOCJOLOGICZNY
Rok XLII — zeszyt 2 — 1980

WANDA MARIA GACZEK, MAŁGORZATA HELPA,
ALEKSANDRA KASPRZYK

NIEHIERARCHICZNA ANALIZA SKUPIEŃ — NOWA METODA KLASYFIKACJI ZJAWISK SPOŁECZNO-GOSPODARCZYCH

1. WPROWADZENIE

Celem artykułu jest przedstawienie podstawowych założeń i zasad stosowania niehierarchicznej analizy skupień oraz wskazanie możliwości jej wykorzystania jako metody klasyfikacji, a w szczególności metody badania struktury przestrzennej zjawisk społeczno-ekonomicznych. Przykład zastosowania tej metody w analizie społeczno-ekonomicznej struktury wsi wielkopolskiej porównuje się z wynikami uzyskanymi dla tych samych danych z diagramu Czekanowskiego. Pozwala to na wykazanie zalet i wad oraz ocenę przydatności tych metod w badaniach przestrzenno-ekonomicznych.

Klasyfikacja stanowi jedną z podstawowych metod badawczych, mających na celu identyfikację grup jednostek podobnych do siebie z punktu widzenia jednej bądź wielu cech jednocześnie. Wobec różnorodności technik, a także samych kryteriów wydzielania klas, ciągle jednak jest to metoda skomplikowana i niejednoznaczna. Wynika to głównie z braku ogólnie przyjętych, a przede wszystkim precyzyjnych definicji podstawowych pojęć takich, jak: podobieństwo, typ, klasa. Przyczyną niejednoznaczności jest także subiektywizm niektórych etapów postępowania w powszechnie stosowanych technikach klasyfikacji, przy jednoczesnym braku — poza oceną wyczerpywalności i rozłączności — obiektywnych metod oceny poprawności rozwiązań. Niehierarchiczna analiza skupień pozwala wyeliminować przynajmniej niektóre z tych mankamentów.

Identyfikacja klas możliwa jest na drodze podziału zbioru bądź grupowania poszczególnych jego elementów, przy czym zarówno pierwszy, jak i drugi sposób może mieć postać hierarchiczną albo niehierarchiczna.

Techniki grupowania, do których należy analiza skupień, polegają na identyfikowaniu klas w procesie aglomerowania podobnych do siebie jednostek. Umożliwiają one jednoczesne rozpatrywanie wielu cech opi-

sujących elementy danego zbioru. Natomiast w przypadku technik podziału celem jest raczej rozdział zbioru na części etapami, najczęściej jednak na każdym z etapów stosowane jest odmienne kryterium (np. wybrana cecha). Możliwość wykorzystania wielozmiennej analizy statystycznej jest szczególnie istotna w przypadku identyfikacji rzeczywistej struktury takich zbiorów, których elementy nie dają się opisać pojedynczymi, prostymi cechami. Zaletą technik grupowania jest także ich większa prostota i szybkość wykonywania obliczeń, podczas gdy pracochłonność technik podziału jest znacznie większa, szczególnie wtedy, kiedy liczba elementów zbioru przekracza 100 jednostek [1].

Grupowanie typu hierarchicznego zakłada, że każda klasa na poziomie (i) jest częścią większej klasy na poziomie (i + 1), a wszystkie one będą mogły być zsumowane na poziomie *(n — l).* Ten sposób identyfikowania klas jest jednak prawidłowy tylko dla niektórych przypadków — na przykład postać hierarchiczną ma każda struktura typu ewolucyjnego. Jednak w sytuacji, kiedy badany zbiór jednostek nie posiada struktury hierarchicznej, klasyfikacja z takim założeniem może prowadzić do znacznych zniekształceń obrazu rzeczywistości [2].

Wykorzystanie technik grupowania hierarchicznego wymaga przyjęcia wstępnych założeń o charakterze struktury badanego zbioru jednostek. Ich zastosowanie wskazane jest więc głównie dla celów testowania hipotez. Odwrotnie natomiast, techniki grupowania niehierarchicznego nie wymagają przyjęcia wstępnych założeń o strukturze zbioru, a tym samym mogą być wykorzystywane w przypadku całkowitej nieznajomości lub tylko częściowego rozpoznania tej struktury, a więc również jako metody generowania hipotez.

Przedstawione powyżej ogólne uwagi o klasyfikacji pozwalają stwierdzić, że: 1) ze względu na mniejszą pracochłonność i możliwość wielozmiennej analizy techniki grupowania zyskują przewagę nad technikami podziału i 2) grupowanie niehierchiczne ma większy zakres zastosowania niż grupowanie hierarchiczne.

2. NIEHIERARCHICZNA ANALIZA SKUPIEŃ

Poddawany klasyfikacji zbiór jednostek traktuje się w analizie skupień jako „chmurę" punktów wielowymiarowej przestrzeni, których współrzędnymi są wartości zmiennych opisujących te jednostki. Skupieniami będą te części przestrzeni, które posiadają relatywnie dużą gęstość punktów. Punkty, które są położone w przestrzeni wielowymiarowej obok

[1] P. M. Mather, *Computational Methods of Multivariate Analysis in Physical Geography,* London 1976, s. 311.

[2] B. S. Everitt, *Cluster analysis,* w: *The Analysis of Survey Data,* t. 1, *Exploring Data Structure,* red. C. A. O'Muircheartaigh, C. Payne, London 1977, s. 67.

siebie, mają zbliżone współrzędne cech i są do siebie podobne, tworzą skupienie punktów czyli klasę jednostek. Przy tak określonej definicji skupienia problemem staje się wskazanie jego centrum, a następnie określenie granic czyli wyznaczenie zasięgu skupienia. Wyróżnia się więc tutaj dwie podstawowe fazy postępowania: 1) identyfikację rdzeni skupień i 2) alokację pozostałych jednostek do tych rdzeni[3]. W aktualnych opracowaniach fazę pierwszą rozbudowuje się o procedury określania liczby klas (czyli skupień) lub wskazania maksymalnej i minimalnej ich liczby, natomiast w końcowym etapie przeprowadza się statystyczną ocenę poprawności rozwiązania.

Centra skupień mogą być wybrane arbitralnie, na podstawie przeglądu macierzy obserwacji, gdzie elementami są wartości cech poszczególnych jednostek[4]. Ustalone rdzenie skupień stanowić będą pewne jednostki typowe dla danego zbioru. Postępowanie takie jest jednak subiektywne, a ponadto może być bardzo trudne lub wręcz niemożliwe w przypadku dużej liczby elementów i cech. Inną drogą określania „początkowych" centrów skupień jest zastosowanie generatora liczb pseudolosowych z przedziału (0, 1) i wybór rdzeni z wartości zawartych pomiędzy maksimum a minimum każdej obserwowanej cechy. Tak ustalone centra skupień nie muszą być konkretnymi jednostkami. Wymaga się tutaj dodatkowo podania największej i najmniejszej dopuszczalnej liczby skupień. Maksymalna liczba skupień musi być tak dobrana, aby była większa od wstępnie przewidywanej, co pozwoli na ewentualną korektę współrzędnych. centrów w dalszych etapach postępowania. Ten drugi sposób wyboru centrów skupień wydaje się być metodą bardziej obiektywną, a jednocześnie mniej pracochłonną od poprzedniej. Jest to ponadto jedyna możliwa droga postępowania przy dużej liczbie jednostek i cech.

Alokacja poszczególnych jednostek do określonego skupienia przeprowadzana jest na podstawie ich odległości od centrum. Jeżeli pierwotne rdzenie skupień zostały wybrane losowo, istnieje możliwość wielokrotnej korekty zasięgu skupień w zależności od przyjętej liczby klas.

Najczęściej stosowaną miarą podobieństwa jednostek (w tym konkretnym przypadku badamy podobieństwo elementów do, określonych

[3] Z. Chojnicki, T. Czyż, *Metody taksonomii numerycznej w regionalizacji geograficznej,* Warszawa 1973, s. 63.

[4] Cechy mogą być wyrażone w formie zmiennych pierwotnych, zmiennych znormalizowanych, odchyleń od średniej, wartości czynników wspólnych albo składowych głównych. Analizę czynników wspólnych lub składowych głównych wykorzystuje się we wstępnej fazie klasyfikacji w celu zredukowania dużej liczby cech oraz zabezpieczenia warunku ich ortogonalności, co umożliwia wykorzystanie prostej miary odległości w dalszych etapach postępowania. Por. W. M. Garczek, *Problemy wykorzystania analizy czynnikowej w badaniach przestrzenno-ekonomicznych,* w: *Prace z zakresu gospodarki przestrzennej,* Zeszyty Naukowe, seria I, z. 75, Poznań 1978, s. 5 - 23.

w poprzednim etapie postępowania, centrów skupień) jest odległość euklidesowa d_{ij} obliczona na podstawie wzoru:

$$d_{ij}=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})\right]^{\frac{1}{2}} \quad i,j=1,2,\ldots,n, \tag{1}$$

gdzie: x_{ik}, x_{jk} — wartości zmiennej k dla obiektu i-tego i j-tego.

Wartości d_{ij} wahają się w granicach od 0,0 (zupełne podobieństwo) do nieskończoności (całkowite niepodobieństwo). Ponieważ jednak wzrastają one wraz z liczbą cech przyjętych do analizy, najczęściej stosuje się normalizowanie d_{ij}, czyli dzielenie jej przez liczbę cech *(p)* otrzymując w efekcie współczynnik odległości. Wykorzystanie tej miary podobieństwa możliwe jest jednak wyłącznie wtedy, kiedy przyjęte do analizy cechy są ortogonalne. Natomiast w przypadku, kiedy zmienne będące współrzędnymi jednostek zbioru są skorelowane albo kiedy dokonujemy klasyfikacji na podstawie wartości czynników rotowanych ukośnie, podobieństwo jednostek określić można na podstawie wzoru uwzględniającego współczynnik korelacji między x_k i x_l [5]:

$$d_{ij}=\left[\sum_{k=1}^{p}\sum_{l=1}^{p}(x_{ik}-x_{jk})(x_{il}-x_{jl})r_{kl}\right]^{\frac{1}{2}} \quad i,j=1,2,\ldots,n. \tag{2}$$

Arbitralny wybór rdzeni skupień determinuje w pewnym stopniu ich granice. Jednostki zostają przydzielone do takiego skupienia, gdzie kwadrat odległości d_{ij} między daną jednostką a tą, którą wybrano jako centrum jest najmniejszy. W konsekwencji zbiór elementów zostaje jednoznacznie podzielony na tyle klas, ile centrów ustalono na wstępie.

Przy losowym wyborze rdzeni skupień ich granice ustalone są w procesie iteracyjnym. Wstępnie poszczególne jednostki zostają również przydzielone do określonych skupień na podstawie najmniejszych odległości cd centrów. W dalszym ciągu ponownie określa się współrzędne centrum jako średnie tych wszystkich jednostek, które zostały wstępnie przyporządkowane do danego skupienia. Powoduje to konieczność sprawdzenia i ewentualnej korekty przynależności jednostek do skupień (mogą ulec zmianie odległości d_{ij}). Ten proces iteracyjny prowadzony jest tak długo, jak długo występuje możliwość przesunięć przy danej liczbie klas. Jeżeli liczba skupień, określona w sposób przedstawiony powyżej, jest większa od założonej na wstępie minimalnej liczby klas, przeprowadza się ich redukcję. Zostają połączone ze sobą te skupienia, dla których kwadrat odległości d_{ij} między centrami jest najmniejszy. Powtórnie określa się również współrzędne dla centrum nowo określonego skupienia, a następnie przeprowadza korektę przynależności jednostek oraz współrzędnych centrów.

Końcowe wyniki analizy umożliwiają dokonanie stosunkowo szczegó-

[5] Według P. M. Mather, op. cit., s. 314.

łowej charakterystyki wyodrębnionych klas. Dla każdej klasy otrzymujemy dane obrazujące współrzędne centrum (w kategoriach cech wyjściowych, np. wartości składowych głównych), liczbę należących jednostek, odległości poszczególnych jednostek oraz ich średnią odległość od centrum. Informacje te pozwalają określić nazwę danego skupienia oraz stopień jego zwartości czy rozproszenia, niezależnie od tego czy jest to wiązka cech, czy klasa jednostek przestrzennych. Subiektywną decyzją w przedstawionej metodzie klasyfikacji jest wybór liczby skupień, a tym samym określenie liczby klas dla danego zbioru jednostek. Analogicznie, w analizie czynników wspólnych decyzją taką jest wybór liczby czynników istotnych, chociaż w tym przypadku istnieją większe możliwości uzasadnienia — na przykład kryterium Kaisera.

W analizie skupień stosuje się dwie metody określania „optymalnej" liczby klas. Mogą one być oczywiście wykorzystywane w przypadku, kiedy liczba ta nie jest ściśle zdeterminowana na wstępie, czyli jedynie przy losowym wyborze icentrów. Metodą najprostszą jest zastosowanie tekstu F dla hipotezy zerowej, że mniejsza liczba skupień nie poprawi rozwiązania. Wyniki końcowe analizy doprowadza się do kilku wersji rozwiązań z różną liczbą skupień i ich pełną charakterystyką. Wersje te sprawdza się jednym z testów statystycznych oraz ocenia w kategoriach możliwości interpretacyjnych lub możliwości przewidywania. Dokonany w ten sposób wybór „optymalnej" liczby skupień jest jedynie przybliżeniem do rzeczywistej struktury zbioru jednostek. Oparty jest on głównie na przesłankach statystycznych, a konkretnie — w przypadku testu F — na minimalizacji wartości odległości jednostek od centrum, jakie wystąpiły przy danej liczbie skupień. Jest to więc ocena raczej stopnia skupienia jednosltek przy określonej liczbie klas, a nie odwzorowania rzeczywistych współzależności między cechami jednostek. Dlatego też wykorzystanie testów statystycznych jako metody określania liczby skupień musi być stosowane z pewną ostrożnością [6].

Druga z metod — nazywana techniką nieliniowego mapowania — daje możliwość wglądu w strukturę zbioru, a tym samym określenie rzeczywistej liczby klas. Jest ona bardziej skomplikowana w sensie obliczeniowym old poprzedniej. Technika nieliniowego mapowania umożliwia graficzne przedstawienie jednostek opisanych współrzędnymi w przestrzeni p-wymiarowej w konfiguracji niskowymiarowej przestrzeni p^* [7], gdzie

[6] B. S. Everitt, op. cit., s. 73.

[7] Metody graficznego przedstawiania zbioru jednostek w przestrzeni dwuwymiarowej stosowane są zarówno w grupowaniu hierarchicznym, np. dendryt najkrótszych odległości Mahalanobisa (Z. Kaczmarek, J. J. Parysek, *Zastosowanie analizy wielowymiarowej w badaniach geograficzno-ekonomicznych,* w: *Metody ilościowe i modele w geografii,* red. Z. Chojnicki, Warszawa 1977, ss. 94--127), jak i grupowaniu niehierarchicznym (P. M. Mather, op. cit., s. 331 - 380; B. S. Everitt, *Graphical Techniques for Multivariate Data,* London 1978, s. 5 - 94).

$p^*<p$, najczęściej $p^*=2$. Dokonuje się tego poprzez minimalizację wyrażenia [8]:

$$E=\frac{1}{\sum_{i<j} d_{ij}}\sum\frac{(d_{ij}-d_{ij}^*)^2}{d_{ij}}, \tag{3}$$

gdzie: d_{ij} jest odległością między jednostkami i i j w pierwotnej p-wymiarowej przestrzeni, d_{ij}^* jest odległością między tymi jednostkami w przestrzeni p^*-wymiarowej.

Jeżeli $p^*=2$, jako podstawę nowej konfiguracji najczęściej wykorzystuje się wartości dwóch pierwotnych zmiennych z maksymalną wariancją, a następnie dokonuje skalowania układu punktów tak, aby E było minimalne. Wartość E jest ważoną sumą kwadratów odległości. Przyjęte wagi dążą do zabezpieczenia „lokalnej", pierwotnej struktury zbioru w tym sensie, że każdy punkt w p^*-wymiarowej przestrzeni będzie dążyć do wykazania tych samych współzależności między sąsiednimi punktami co w przestrzeni p-wymiarowej. Jego współzależności z punktami oddalonymi (czyli odmiennymi) w przestrzeni pierwotnej będą również odmienne w tym samym zakresie w przestrzeni p^*.

Graficzny obraz układu jednostek zbioru w przestrzeni dwuwymiarowej pozwala w przybliżeniu ustalić liczbę skupień. Ponieważ jednak same odległości d_{ij}^* mogą być odmienne od d_{ij} (przy zachowaniu współzależności między jednostkami!), nie można za pomocą tej techniki ustalić dokładnych granic skupienia, czyli przynależności jednostek do danej klasy. Dlatego też nieliniowe mapowanie wykorzystywane we wstępnej fazie analizy skupień jest metodą identyfikacji rzeczywistej liczby klas, natomiast stosowanie tej techniki w końcowym etapie analizy będzie jedynie techniką kartograficzną służącą do graficznego przedstawiania układu danych w uproszczonej wersji.

Podobnie jak w każdej ze stosowanych dotychczas metod klasyfikacji, również w analizie skupień niektóre z decyzji mają charakter arbitralny i obarczają wyniki pewną dozą subiektywizmu. Należą do nich: dobór cech opisujących jednostki zbioru, identyfikacja centrów i ustalenie liczby skupień (klas), a także, przy pewnych celach badawczych, nazywanie (tzw. etykietowanie) skupień. Część z tych decyzji wpływa ma stopień subiektywizmu każdej z metod klasyfikacji, specyficznym elementem prezentowanej metody jest natomiast wybór liczby skupień. Z tego powodu często stosuje się równolegle obydwie przedstawione techniki ustalania ich liczby: nieliniowe mapowanie na wstępie i test statystyczny w końcowym etapie. Jednakże nawet przy takim postępowaniu wskazane jest, aby skupienia oceniane były również na podstawie wyników uzyskanych z innych metod klasyfikacji i możliwości interpretacji klas.

[8] B. S. Everitt, op. cit., s. 30.

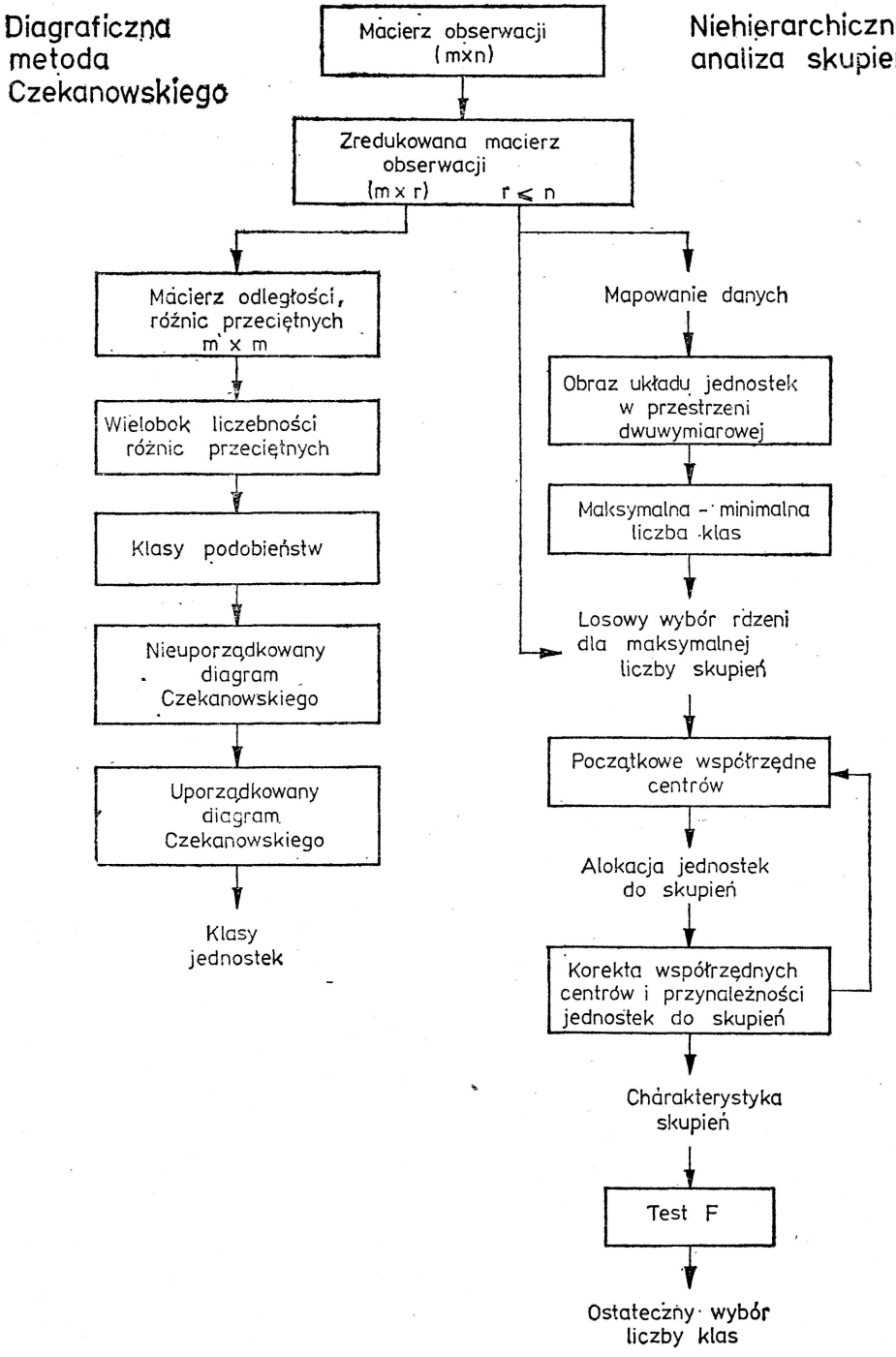
3. PRZYKŁAD OKREŚLANIA TYPÓW STRUKTURY PRZESTRZENNEJ ZJAWISK SPOŁECZNO-EKONOMICZNYCH

Niehierarchiczna analiza skupień jest metodą, którą można wykorzystać dla celów typologii bądź rejonizacji różnorodnych zjawisk. Mogą to być badania dotyczące określania rejonów poziomu uprzemysłowienia i urbanizacji terenu czy charakterystyki społeczno-demograficznej. Jednocześnie metoda może służyć wykrywaniu charakterystycznych typów przestrzennej struktury zjawisk społeczno-ekonomicznych istotnych ze względu na cel badań, czyli na przykład typów produkcji rolnej wsi, typów funkcjonalnych miast, typów sieci transportowej itd.

Typy przestrzennej struktury społeczno-ekonomicznej wsi wielkopolskiej określono na podstawie dwóch metod: diagraficznej J. Czekanowskiego i niehierarchicznej analizy skupień. Postępowanie takie pozwalało zrealizować cel metodologiczny pracy: porównać wyniki każdej z metod, wskazać ich zalety i wady, a przede wszystkim ocenić niehierarchiczna analizę skupień na tle tradycyjnie wykorzystywanego w literaturze polskiej dla celów klasyfikacji diagramu Czekanowskiego. Obok tego podjęto również cel poznawczy, którym było obiektywne rozpoznanie typów struktury, wykazanie stałości i uniezależnienia od stosowanej metody, gdyż jedynie takie typy mogą być punktem wyjścia dla badań przestrzennej struktury zjawisk. Etapy postępowania każdej z wykorzystywanych metod przedstawiono na rycinie 1. Początkowe fazy, niejako kroki wstępne, są takie same dla każdej z nich.

Jako jednostki podstawowe przyjęto 29 powiatów byłego województwa poznańskiego. Nie są one jednorodne wewnętrznie, jednakże zebranie danych w mniejszej skali przestrzennej było bardzo utrudnione, a dla niektórych charakterystyk wręcz niemożliwe. Jednostki te opisano 60 zmiennymi, które odzwierciedlały społeczno-ekonomiczny charakter wsi. Były to cechy opisujące: 1) poziom gospodarki rolnej (w tym warunki naturalne rolnictwa, strukturę agrarną, poziom mechanizacji i nawożenia, poziom produkcji rolnej — razem 27 cech), 2) strukturę ludności wiejskiej (w tym: strukturę zawodową, strukturę demograficzną, strukturę wykształcenia — razem 23 cechy), poziom uprzemysłowienia i 4) poziom urbanizacji powiatów oraz 5) warunki mieszkaniowe ludności wiejskiej.

Macierz obserwacji, kończąca pierwszy etap postępowania miała wymiary 29×60. Względy ekonomiczne (głównie czas obliczeń), a także chęć określenia syntetycznych wymiarów struktury społeczno-ekonomicznej wsi przemawiały za zredukowaniem liczby cech. Jako metodę redukcji przestrzeni wielocechowej wybrano analizę czynników wspólnych, która uważana jest za najbardziej efektywną. Umożliwia ona także ortogonalizację cech opasujących zbiór jednostek, a tym samym wykorzystanie prostej miary odległości euklidesowej jako miary podobieństwa.


Diagraficzna metoda Czekanowskiego
Niehierarchiczna analiza skupień
Macierz obserwacji (m×n)
Zredukowana macierz obserwacji (m x r) r ≤ n
Macierz odległości, różnic przeciętnych m x m
Wielobok liczebności różnic przeciętnych
Klasy podobieństw
Nieuporządkowany diagram Czekanowskiego
Uporządkowany diagram Czekanowskiego
Klasy jednostek
Mapowanie danych
Obraz układu jednostek w przestrzeni dwuwymiarowej
Maksymalna - minimalna liczba klas
Losowy wybór rdzeni dla maksymalnej liczby skupień
Początkowe współrzędne centrów
Alokacja jednostek do skupień
Korekta współrzędnych centrów i przynależności jednostek do skupień
Charakterystyka skupień
Test F
Ostateczny wybór liczby klas

Ryc. 1.

Analiza czynników wspólnych[9] pozwoliła wyodrębnić pięć podstawowych wymiarów przestrzeni będącej przedmiotem niniejszych badań.

Wymiar FI, interpretowany jako poziom struktury agrarnej oraz natężenie zatrudnienia rolniczego, wyjaśniał 28,05% wariancji wspólnej. Jednostki o wysokich dodatnich wartościach FI to powiaty o przewadze dużych gospodarstw indywidualnych (ponad 10 ha) i znacznym udziale sektora uspołecznionego, a jednocześnie małym natężeniu zatrudnienia w rolnictwie. Natomiast jednostki o ujemnych wartościach FI to powiaty o przewadze średnich gospodarstw rolnych, niskiem uspołecznieniu rolnictwa <oraz dużej liczbie zatrudnionych w rolnictwie na 100 ha użytków rolnych.

Wymiar F2 zidentyfikowany został jako poziom intensywności rolnictwa (27,91% wariancji wspólnej). Jednostki z wysokimi dodatnimi wartościami tego czynnika to powiaty z przewagą gleb lepszych niż przeciętne, o najwyższym w Wielkopolsce poziomie produkcji rolnej (roślinnej i zwierzęcej), wysokim poziomie towarowości i znacznym poziomie mechanizacji i nawożenia w gospodarstwach indywidualnych.

Wymiar F3 określony został jako poziom urbanizacji ekonomicznej wsi (19,43% wariancji wspólnej). Charakteryzuje on natężenie procesu suburbanizacji — wysokie dodatnie wartości F3 są charakterystyczne dla powiatów o znacznym udziale zatrudnionych poza rolnictwem wśród czynnej zawodowo ludności wiejskiej (np. powiat Poznań).

Wymiar F4 zidentyfikowany został jako charakter stosunków pracy i typ produkcji rolnej (10,60% wariancji wspólnej). Interpretacja wartości czynników w jednostkach jest w przypadku tego wymiaru utrudniona. Ogólnie wysokie dodatnie wartości F4 charakterystyczne są dla jednostek z dużym udziałem ludności dwuzawodowej na wsi, przy czym podstawą utrzymania są zajęcia pozarolnicze. Natomiast ujemne wartości F4 charakteryzują powiaty z największym udziałem gospodarstw indywidualnych zatrudniających pracowników najemnych (gospodarstwa specjalistyczne). Równocześnie wymiar ten pozwala charakteryzować typ produkcji rolnej.

Wymiar F5 zidentyfikowany został jako poziom kolektywizacji (5,64% wariancja wspólnej). Jednostki o maksymalnych wartościach F5 to powiaty o najwyższym udziale rolniczych spółdzielni produkcyjnych w ogólnym areale gruntów, a także najwyższym udziale za-

[9] Czynniki wspólne obliczono metodą największej wiarygodności *(ML)* na podstawie macierzy korelacji; liczbę istotnych czynników przyjęto zgodnie z kryterium Kaisera (wartość własna>1). Osie czynników przed interpretacją były rotowane ortogonalnie według kryterium Varimax. Obliczenie i adaptację programów wykonano w Ośrodku Przetwarzania Informacji Akademii Ekonomicznej w Poznaniu.

Tabela 1

Wartości czynników w jednostkach

Numer jednostki	Powiat	*F1*	*F2*	*F3*	*F4*	*F5*
1	Chodzież	16,3499	−5,5939	6,2565	0,3147	1,9026
2	Czarnków	0,5220	−16,9668	0,9998	8,8361	−2,6438
3	Gniezno	13,6713	10,3064	−3,6311	−5,4818	1,4011
4	Gostyń	−1,1637	30,6113	−3,2543	2,4905	2,2530
5	Jarocin	−2,9049	9,5059	13,4684	4,0823	2,6289
6	Kalisz	−25,5760	−8,9332	−10,1092	−6,8989	−7,8809
7	Kępno	−8,6385	−6,6186	2,1865	7,6751	−2,6622
8	Koło	−27,9579	−18,2256	−18,5950	−5,7025	−8,6708
9	Konin	−23,4375	−20,6631	−5,4563	−4,8536	−6,6198
10	Kościan	1,9624	12,0233	−3,1638	−0,9259	5,2081
11	Krotoszyn	−2,1066	20,7070	−10,3288	−0,6662	0,5611
12	Leszno	11,3298	11,6968	9,7853	4,1986	3,3887
13	Międzychód	20,5154	−1,9299	0,9852	−1,0862	1,4382
14	Nowy Tomyśl	6,8185	−1,3170	4,3829	3,6141	0,8689
15	Oborniki	14,3238	4,5576	2,9597	−4,8576	7,0194
16	Ostrów Wlkp.	−7,5040	−9,2317	13,4657	8,6630	−1,7652
17	Ostrzeszów	−19,1201	−19,6153	−9,1192	8,0364	−6,7583
18	Pleszew	−9,0252	0,2352	−2,5997	1,1792	−1,5667
19	Poznań	17,2292	5,0495	33,8293	−9,7638	5,2936
20	Rawicz	−4,8946	18,8835	−6,9005	2,5521	−1,2462
21	Słupca	−18,3154	−11,9647	−15,3304	−4,3099	−6,5648
22	Szamotuły	16,5752	16,3518	7,7410	−3,1012	7,9685
23	Śrem	11,8438	10,2639	5,1530	−1,8204	8,2386
24	Środa	15,4460	17,3215	3,6957	−6,0246	8,9437
25	Trzcianka	16,0694	−22,9134	−2,0476	3,1917	−3,5406
26	Turek	−28,7275	−20,6601	−18,1761	−4,5522	−8,4123
27	Wągrowiec	15,6717	4,2459	−5,3045	−2,7563	0,8580
28	Wolsztyn	−4,7407	−15,6205	3,6808	10,9415	−0,6278
29	Września	5,7843	8,4942	5,4268	−3,6943	0,9861

trudnionych w RSP w ogólnej liczbie czynnyich zawodowo na wsi. Wartości czynników w poszczególnych jednostkach przedstawia tabela 1. Stanowią one podstawę klasyfikacja zbioru oraz ustalania typów struktury społeczno-ekonomicznej wsi wielkopolskiej.

Metoda Czekanowskiego [10] pozwala na wyodrębnienie klas na podstawie tzw. uporządkowanego diagramu. Nieuporządkowany diagram Czekanowskiego jest właściwie graficznym obrazem macierzy odległości,

[10] Matematyczne i techniczne zasady stosowania metody Czekanowskiego w badaniach przestrzenno-ekonomicznych przedstawione są m. in. w pracach: B. Podolec, K. Zając, *EkonoTnetryczne metody ustalania rejonów konsumpcji,* Warszawa 1978; Z. Chojnicki, T. Czyż, *Metody taksonomii numerycznej w regio¬ nalizacji geograficznej,* Warszawa 1973.

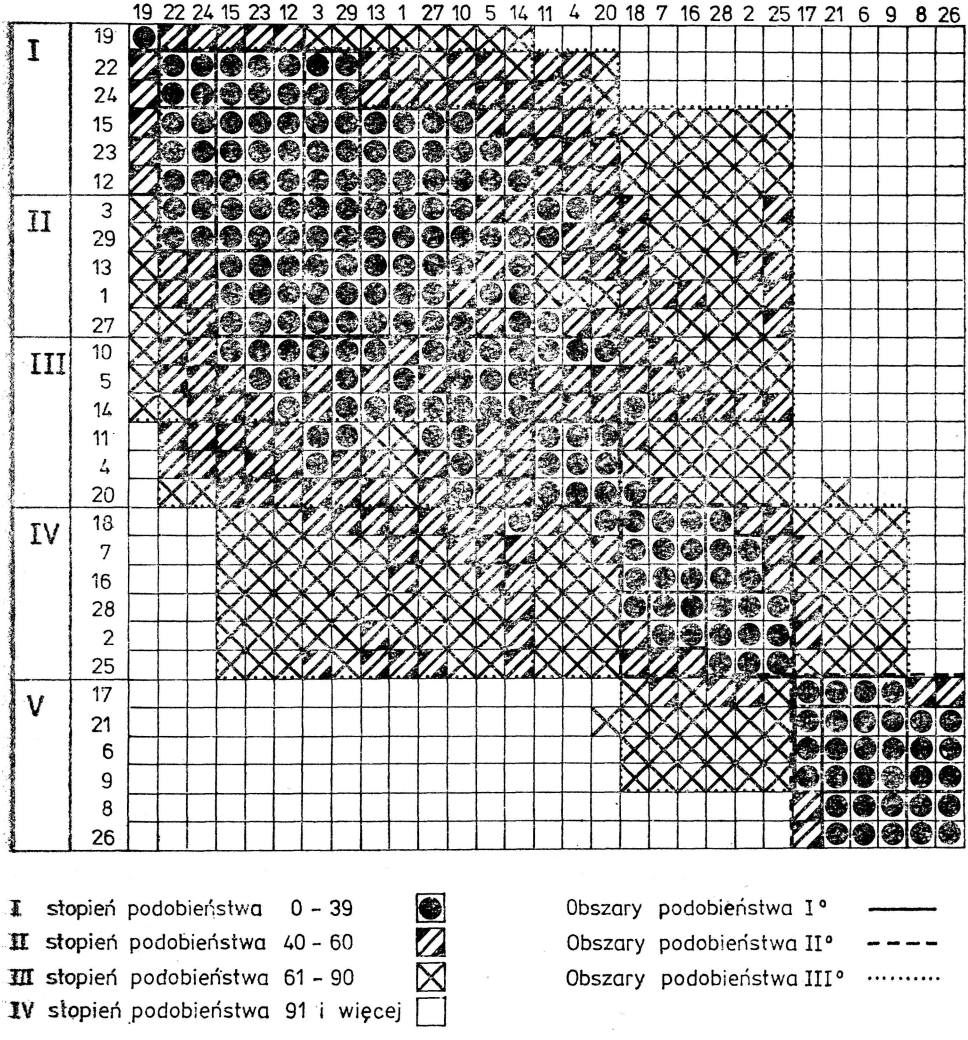
której elementami są odległości pomiędzy każdą parą jednostek ze względu na każdą z cech (tutaj są to odległości między każdą parą 29 jednostek ze względu na 5 czynników). Jego wykreślenie wymaga ustalenia skali podobieństw, przy czym poprawność tej skali warunkuje obiektywne ustalenie klas, w ramach których jednostki są bardziej do siebie podobne niż między klasami. W tym przykładzie klasy podobieństw wyznaczono na podstawie wieloboku liczebności sumarycznych różnic. Uporządkowanie diagramu sprowadza się do takiego przestawienia rzędów i kolumn pierwotnego układu, aby elementy najbardziej podobne wyodrębniły się jak najwyraźniej wzdłuż przekątnej. Porządkowanie diagramu jest czynnością bardzo pracochłonną szczególnie wtedy, kiedy liczba jednostek jest znaczna. Często wykorzystuje się na tym etapie dodatkowo dendryt, który ułatwia ustalenie kolejności przestawienia rzędów i kolumn diagramu.

Dendryt, czyli graficzne uporządkowanie zbioru tak, aby pary elementów o najkrótszych odległościach (najbardziej podobne do siebie) sąsiadowały ze sobą, jest samoistnie również wykorzystywany jako metoda klasyfikacji. Umożliwia on jednak przedstawienie na płaszczyźnie jedynie najmniejszych odległości między obiektami, abstrahując od wszystkich pozostałych. Tym samym nie daje pełnego obrazu przestrzeni. Brak również statystycznego kryterium ustalenia optymalnej wartości granicznej d_{ij} (r). Wnioskowanie o strukturze zbioru na podstawie podziału dendrytu jest tym samym ograniczone. Zaletą tej metody jest, jednakże jej duża czytelność i poprawność formalna procedury.

Analiza uporządkowanego diagramu Czekanowskiego (ryc. 2) wykazała istnienie pięciu typów struktury społeczno-ekonomicznej wsi wielkopolskiej w 1970 r. Przy wyodrębnianiu podzbiorów jednostek uwzględniono podobieństwa 1 i 2 stopnia. Wydzielenie takich właśnie podzbiorów jest jednak do pewnego stopnia arbitralne. Jednostkę 19 (powiat Poznań), przy uwzględnieniu 2 stopnia podobieństwa można włączyć do typu I, jednakże może ona być także traktowana jako samodzielna grupa jednoelementowa. Również określenie granic typu II i III jest utrudnione, może należałoby ostatni z nich rozbić na dwa podtypy. Zdecydowanie natomiast wyodrębniają się w diagramie jednostki tworzące typ IV i V; ich odrębność była już wyraźnie widoczna na dendrycie, który tutaj wykorzystano jako pomoc przy porządkowaniu diagramu.

Wydzielenie klas na podstawie uporządkowanego diagramu Czekanowskiego uzależnione jest faktycznie od wizualnego przeglądu tablicy odległości[11]. Jest cno łatwe i nie wzbudza kontrowersji, jeżeli w zbiorze

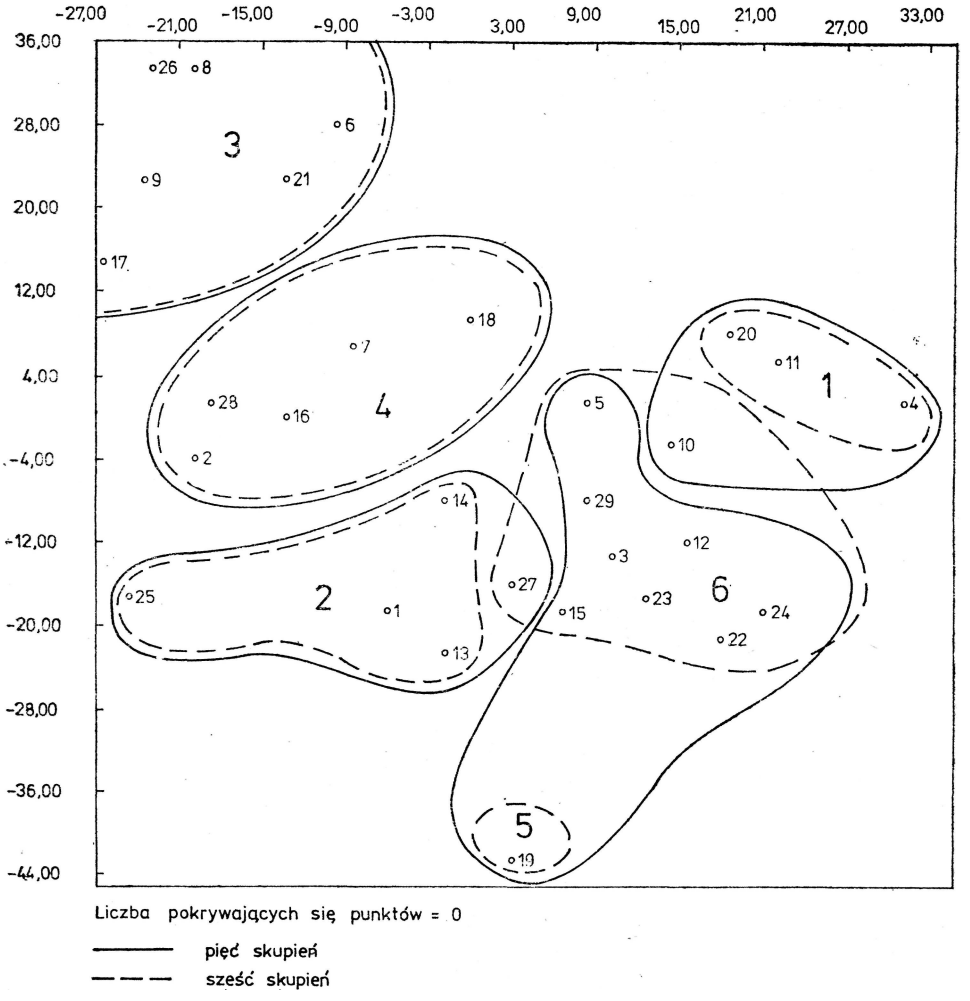
[11] Bazując na idei Czekanowskiego opracowano nowe metody taksonomiczne: numeryczną metodę SHADE opartą na macierzy odległości Mahalanobisa, która daje w wyniku graficzny obraz nieuporządkowanego i uporządkowanego diagramu podobieństwa (R. F. Ling, *A computer generated aid for cluster analysis,* Communications of the AMC, 1973, nr 16, s. 355 - 361) oraz metodę ORLINE



Ryc. 2

wyodrębniają się zdecydowanie podgrupy jednostek. Sytuacja taka jest jednak mało prawdopodobna, kiedy cechy charakteryzujące elementy obioru opisują różnorodne, chociaż oczywiście powiązane ze sobą zjawiska, a struktura zbioru nie jest hierarchiczna. W prezentowanym przykładzie jednostki układały się liniowo, co dodatkowo utrudniało ustalenie liczby typów i ich wielkości. Określenie typów struktury społeczno-

(Z. Piasecki, *Nowe metody taksonomiczne i ich właściwości klasyfikujące,* Sympozjum na temat: *Zastosowanie metod taksonomicznych w geografii,* Poznań 1977). Obydwie te metody, dzięki wykorzystywaniu ilościowego kryterium podobieństwa między obiektami, eliminują w znacznym stopniu subiektywizm metody Czekanowskiego.



Ryc 3

-ekonomicznej wsi za pomocą niehierarchicznej analizy skupień [12] wymagało na wstępie podjęcia decyzji o liczbie skupień i sposobie ustalania centrów. Pomija się tutaj odległości między poszczególnymi elementami zbioru: punktem wyjścia są wartości czynników w jednostkach (tab. 1).

Abstrahując od rozpoznanej już na podstawie diagramu Czekanowskiego struktury zbioru jednostek, jako metodę określenia liczby skupień, wykorzystano technikę nieliniowego mapowania. Układ punktów odpowiadających poszczególnym jednostkom zbioru w zredukowanej przestrzeni dwuwymiarowej przedstawia rycina 3 (linie stanowiące gra-

[12] Program AN/6 (W. M. Gaczek, A. Kasprzyk, A. Chmielewska-Wawrzyniak, *Niehierarchiezna analiza skupisk,* Poznań 1979) oraz obliczenia nieliniowego mapowania i niehierarchicznej analizy skupień wykonano w Ośrodku Przetwarzania Informacji Akademii Ekonomicznej w Poznaniu.

Tabela 2

Test *F* dla wyboru liczebności skupień

Numer skupienia	Liczba skupień	*F*	*DF*1	*DF*2
2	3	1,9235	5	130
3	4	2,3472	5	125
4	5	1,5836	5	120
5	6	2,5406	5	115
6	7	1,6236	5	110
7	8	1,8101	5	105

nice skupień wykreślono po zakończeniu wszystkich obliczeń). Wykres pozwala zorientować się czy w danym zbiorze występują skupienia i jedynie ogólnie określić ich liczbę. Obserwuje się znaczne rozproszenie punktów i dlatego właściwa interpretacja wykresu jest utrudniona. Nie powinno to być jednak zaskoczeniem, gdyż zbiór cech pierwotnie opisujących jednostki był bardzo obszerny i zróżnicowany. Uzyskany układ punktów świadczyć może także o niehierarchicznej strukturze zbioru. Na podstawie otrzymanego układu punktów przyjęto wstępnie założenie, że zbiór jednostek wykazuje pięć lub sześć skupień. Decyzję ostateczną uzależnia się od testu *F*.

Początkowe centra skupień określono w sposób losowy — progran zakładał takie obliczenie współrzędnych rdzeni, laby znajdowały się między maksymalną a minimalną wartością każdego z czynników, czyli nie musiały to być konkretne jednostki. Przedział minimalnej i maksymalnej liczby skupień został rozszerzony w stosunku do ustalonej na podstawie obrazu układu punktów w uproszczonej przestrzeni dwuwymiarowej: liczba maksymalna = 8, minimalna = 3. Zapewnia to możliwość ewentualnej korekty granic skupień oraz umożliwia dokonanie statystycznej oceny poprawności rozwiązania, czyli zwiększa obiektywność metody.

Wyniki otrzymane w końcowym etapie zawierały tabulogramy z ogólną i szczegółową charakterystyką klas dla każdej wersji liczebności skupień oraz test *F,* oceniający, która z wersji jest optymalna w sensie statystycznym. Wyznaczone wartości testu *F* dla oceny Ho: mniejsza liczba skupień poprawia rozwiązanie, zawiera tabela 2. Na poziomie ufności 0,05 wartość statystyki *F* dla wyznaczonych stopni swobody wynosi 4,40 [13], co nie pozwala odrzucić hipotezy zerowej. Wartością najbardziej zbliżoną do wartości granicznej statystyki *F* jest rozwiązanie z 6 skupieniami i ta wersja została przyjęta jako ostateczna. Rozwiązanie takie jest również zgodne z obrazem układu jednostek w przestrzeni dwuwymiarowej (ryc. 3).

[13] J. E. Freund, *Podstawy nowoczesnej statystyki,* Warszawa 1971, Tabl. IVa.

Tabela 3

Ogólna charakterystyka wyodrębnionych typów (dla 6 skupień)

Numer skupienia	Liczba jednostek	Współrzędne centrów skupień					Średnia odległość jednostki od centrum
		*F*1	*F*2	*F*3	*F*4	*F*5	
1	3	−2,7261	23,4006	−6,8279	1,4587	−0,5226	6,270
2	4	14,9383	−7,9386	2,3943	1,5086	−0,1673	10,270
3	6	−23,8557	−16,6770	−12,7977	−3,0468	7,4845	9,157
4	5	−5,8773	−9,6405	3,5466	7,4590	1,8531	9,084
5	1	17,2292	5,0495	33,8293	−9,7638	−5,2936	0,000
6	10	10,3703	10,4767	3,6130	1,9661	−4,6642	9,934

Tabela 4

Charakterystyka typu 6

Jednostka	Odległość od centrum
3	9,2957
5	18,0071
10	10,9722
12	8,9519
15	8,0623
22	10,1126
23	4,1696
24	10,3637
27	12,7105
29	6,6908

Wyodrębniono sześć typów struktury społeczno-ekonomicznej wsi wielkopolskiej, które scharakteryzować można na podstawie: współrzędnych centrów skupień i średniej odległości jednostek od centrum (tab. 3), odległości poszczególnych jednostek od centrum (tab. 4) i macierzy odległości między centrami skupień (tab. 5).

Typ 1 (jednostki: Gostyń, Krotoszyn, Rawicz). Charakterystyczny dla tego typu jest najwyższy w zbiorze poziom intensywności produkcji rolnej (maksymalne wartości *F2* dla centrum — tab. 4) przy stosunkowo niskim poziomie urbanizacji wsi i średnim poziomie dwuzawodowości. Wśród gospodarstw indywidualnych występuje znaczny udział gospodarstw o dużej powierzchni użytków rolnych, a jednocześnie obserwuje się niewielkie zatrudnienie pracowników najemnych w rolnictwie. Ten typ struktury społeczno-ekonomicznej wsi wykazuje maksymalne skupienie jednostek; są one w zakresie charakterystyki wszystkich wymiarów struktury najbardziej podobne do siebie i tym samym średnia odległość od centrum skupienia jest najniższa wśród wyodrębnionych ty-

pów wieloelementowych (tab. 3). Jednostki tego typu są również zwarte przestrzennie.

Typ 2 (jednostki: Chodzież, Trzcianka, Międzychód, Nowy Tomyśl). Jednostki tego typu wyróżnia przede wszystkim specyficzna struktura agrarna (duży udział sektora państwowego i znaczny procent dużych gospodarstw indywidualnych — ponad 10 ha) przy niskim poziomie intensywności rolnictwa (poniżej średniej w Wielkopolsce). Jednocześnie są to tereny o niskim natężeniu zatrudnienia rolniczego na 100 ha użytków rolnych, o średnim poziomie urbanizacji ekonomicznej wsi *i* dwuzawodowości.

Typ 3 (jednostki: Kalisz, Koło, Konin, Ostrzeszów, Słupca, Turek). Jest to zwarty obszar wschodniej części byłego województwa poznańskiego. Struktura społeczno-ekonomiczna wsi tych terenów jest najbardziej odrębna od przeciętnej w Wielkopolsce. W strukturze agrarnej dominuje sektor indywidualny, poziom uspołecznienia rolnictwa jest zdecydowanie najniższy. Wśród gospodarstw indywidualnych przeważają gospodarstwa średnie i małe (5-10 ha). Poziom intensywności rolnictwa jest najniższy w zbiorze (minimalne wartości *F2*), natomiast natężenie zatrudnienia rolniczego wysokie (maksymalna liczba czynnych zawodowo w rolnictwie na 100 ha użytków rolnych), a wśród ludności dwuzawodowej dominują chłopi-robotnicy czyli osoby, które zatrudnienie poza rolnictwem traktują jedynie jako dodatkowe źródło utrzymania. Ponadto występuje tutaj najniższy poziom suburbanizacji (minimalna wartość *F3* dla centrum — tab. 3).

Występowanie tego typu struktury społeczno-ekonomicznej wsi ma oczywiście uzasadnienie historyczne. Większość tych ziem została przyłączona do województwa poznańskiego dopiero pod koniec dwudziestolecia międzywojennego, poprzednio należała do województwa łódzkiego, a wcześniej do zaboru rosyjskiego. Jest to więc dziedzictwo innej drogi rozwojowej w okresie przed drugą wojną światową. Okazuje się, że mimo wielu przedsięwzięć zmierzających do wyrównania struktury społeczno-ekonomicznej wsi poszczególnych regionów, nadal do roku 1970 występowało wiele cech odmiennych.

Typ 4 (jednostki: Czarnków, Kępno, Ostrów Wlkp., Pleszew, Wolsztyn). Wyróżnia się on największym w Wielkopolsce poziomem dwuzawodowości, ze znaczną przewagą robotników-chłopów, którzy traktują pracę w rolnictwie jedynie jako dodatkowe źródło utrzymania. Poziom uspołecznienia rolnictwa średni, wśród gospodarstw indywidualnych dość duży udział gospodarstw małych,[1] jednocześnie niższy od średniej dla całego zbioru poziom intensywności rolnictwa.

Typ 5 (jedna jednostka: Poznań). Struktura społeczno-ekonomiczna wsi ukształtowała się pod wpływem aglomeracji poznańskiej: najwyższy poziom suburbanizacji (maksymalna wartość *F3* — tab. 3) czyli znaczny

odsetek ludności zawodowo czynnej na wsi pracuje poza rolnictwem oraz specyficzny charakter produkcji rolnej wynikający z zapotrzebowania rynku zbytu w mieście i wynikające z tego większe znaczenie pracy najemnej w rolnictwie. Poziom intensywności rolnictwa nieco powyżej średniej dla całego zbioru, jednocześnie średni poziom dwuzawodowości i przeciętna struktura agrarna (zarówno w zakresie własności, jak i wielkości gospodarstw).

Tabela 5;

Macierz odległości między centrami skupień

	1	2	3	4	5	6
1	0					
2	37,1375	0				
3	46,6146	43,4906	0			
4	35,3687	21,8404	27,9635	0		
5	50,3638	36,2991	67,3980	44,8826	0	
6	21,8248	19,8436	48,2371	28,2843	32,4148	0

Odległości między rdzeniem tego skupienia a pozostałymi są bardzo duże (tab. 5). Przemawia to za wydzieleniem tej jednostki jako odrębnego typu jednoelementowego, a nie włączeniem jej do któregoś z pozostałych typów tak, jak to przyjęto na podstawie diagramu Czekanówskiego. Również wielkości poszczególnych wymiarów struktury (zwłaszcza *F3* i *F4*, ale także *F5*) świadczą o zdecydowanej odrębności tej jednostki od pozostałych.

Typ 6 (jednostki: Gniezno, Jarocin, Kościan, Leszno, Oborniki, Śrem, Środa, Szamotuły, Wągrowiec, Września). Jest to największa grupa jednostek, a jednocześnie wielkości poszczególnych wymiarów są tutaj zbliżone do typowych tradycyjnych charakterystyk rolnictwa Wielkopolski. Charakteryzuje się on wysokim udziałem sektora uspołecznionego w rolnictwie — znaczna jest liczba gospodarstw spółdzielczych i państwowych. Wśród gospodarstw indywidualnych przeważają jednostki duże (ponad 10 ha powierzchni). Jednocześnie obserwuje się wysoki poziom intensywności rolnictwa a średni suburbanizacji i dwuzawodowości. Jednostką najbardziej zbliżoną do centrum skupienia, czyli najbardziej charakterystyczną w tym podzbiorze, jest jednostka 23 — Śrem, natomiast jednostkami najbardziej odbiegającymi od średniej są jednostki 5 i 27 — Jarocin i Wągrowiec (tab. 4).

Uzyskane wyniki z niehierarchicznej analizy skupień, a w szczególności macierz odległości między centrami (tab. 5) pozwalają także wskazać typy najbardziej „oddalone" od pozostałych — będą to: typ 3 — obszary wschodnie i typ 5 — Poznań. Są to najbardziej specyficzne typy struktury społeczno-ekonomicznej wsi wielkopolskiej w 1970 r. Te właś-

nie grupy jednostek najwyraźniej wyodrębniały się w zbiorze; ich odrębność była widoczna w układzie jednostek w przestrzeni dwuwymiarowej (ryc. 3), a w "przypadku elementów typu 3, również w diagramie Czekanowskiego (ryc. 2), a także dendrycie.

Typy struktury społeczno-ekonomicznej wsi wykryte za pomocą analizy skupień są w pewnych granicach ,podobne do klas określonych na podstawie diagramu Czekanowskiego [14]. Zbieżność idealną wykazuje jedynie typ 3, który odpowiada klasie V z diagramu — w jego skład, jak już wspomniano, wchodzą jednostki maksymalnie odbiegające od średnich dla całego obszaru. Typ 4, zawierający pięć jednostek, jest zbliżony do klasy IV, z tym że jednostka 25 została, w porównaniu z diagramem Czekanowskiego, przesunięta do innego typu. Jednostka (Trzcianka) została włączona w analizie skupień do typu 2 (ryc. 3), jednakże posiada ona w tym typie, a także w całym zbiorze, maksymalną odległość od centrum skupienia (16, 18). Również włączenie jej do typu 4 — co byłoby zgodne z klasyfikacją wynikającą z diagramu — nie zmienia sytuacji, nadal odbiega ona znacznie od rdzenia skupienia. Jej odległość przy takim rozwiązaniu nawet wzrasta (18, 27). Tym samym wyniki uzyskane z analizy skupień wydają się bardziej realne niż wyniki z diagramu.

Pozostałe typy wykazują już pewną odrębność. I tak typ 6 (grupa jednostek najbardziej charakterystycznych dla rolnictwa Wielkopolski) to klasy I i II oprócz powiatu Poznań, który w analizie skupień został wyłączony jako odrębny typ jednoelementowy. Umowność granicy między klasą I i II była widoczna w diagramie, natomiast charakterystyka typu 6 (np. niewielkie rozproszenie elementów — średnia odległość jednostki od centrum 9,93) oraz odległości centrum tego skupienia od pozostałych, a także zwartość przestrzenna tych jednostek wydają się sugerować, że i w tym przypadku wyniki analizy skupień lepiej oddają rzeczywisty układ elementów. Dwa ostatnie typy: 1 i 2, tworzyły w diagramie klasę III, w której wyodrębniają się jednak dwie podgrupy. Nielogiczność połączenia jednostek w tej klasie wyraźnie uwidoczniła się na rycinie 3, gdzie ich skupienia są oddzielone od siebie jednostkami należącymi do innego typu.

Porównanie wyników diagraficznej metody J. Czekanowskiego i niehierarchicznej analizy skupień pozwala stwierdzić, że druga z nich jest metodą lepiej odzwierciedlającą rzeczywistą strukturę zbioru jednostek. Otrzymanie typologii identycznej jak w analizie skupień możliwe było jednak również z diagramu Czekanowskiego, jednakże układ jednostek w diagramie, a także umowność ich podziału dopuszczały w tym przypadku wiele wariantów rozwiązania.

[14] Klasy jednostek określonych na podstawie diagramu Czekanowskiego oznaczono cyframi rzymskimi, natomiast typy wynikające z analizy skupień cyframi arabskimi.

4. UWAGI KOŃCOWE

Na podstawie przeprowadzonych badań i przeglądu literatury z zakresu metod klasyfikacji sformułować można następujące uogólnienia:

1) Niehierarchiczna analiza skupień, będąca metodą klasyfikacji opartą na zasadach grupowania jednostek, ma szeroki zakres zastosowania, ponieważ przy jej wykorzystywaniu nie ma potrzeby przyjmowania wstępnych założeń o strukturze badanego zbioru elementów, tak jak to jest w przypadku technik grupowania hierarchicznego, może ona być stosowana nie tylko jako metoda testowania hipotez, ale przede wszystkim jako metoda ich kreowania.

2) Losowy wybór centrów skupień w znacznym stopniu eliminuje subiektywizm klasyfikacji w przypadku stosowania tej metody. Taki sposób wyboru rdzeni jest szczególnie istotny przy wykorzystywaniu metody właśnie dla celów generowania hipotez. Odwrotnie, arbitralny wybór centrów z macierzy obserwacji ma szczególne znaczenie przy wykorzystywaniu metody do celów testowania hipotez.

3) Niehierarchiczna analiza skupień może być wykorzystywana w badaniach struktury zbiorów o bardzo dużej liczbie elementów i znacznej liczbie cech. Zastosowanie natomiast diagraficznej metody Czekanowskiego jest właściwie ograniczone ze względów technicznych do badań zbiorów o niewielkiej liczbie elementów.

4) Zastosowanie niehierarchicznej analizy skupień pozwala uzyskać jednoznaczne wyniki — następujące ścisłe wyznaczenie przynależności jednostek do określonych klas, a statystyczna ocena (test *F)* poprawności klasyfikacji, zdeterminowana faktycznie poziomem ufności, umożliwia jednoznaczny wybór rozwiązania. Prosta wersja diagraficznej metody Czekanowskiego z reguły dopuszczała wiele wariantów podziału zbioru jednostek.

5) Wyniki niehierarchicznej analizy skupień pozwalają dokonać charakterystyki statystycznej klas oraz dodatkowej charakterystyki wewnętrznej struktury zbioru (typy specyficzne, rozproszenie typów, zwartość).

6) Zaletą niehierarchicznej analizy skupień, w stosunku do technik grupowania hierarchicznego, jest możliwość relokacji tych jednostek zbioru, które zostały ewentualnie źle zakwalifikowane w początkowym etapie analizy.

NON-HIERARCHICAL CLUSTER ANALYSIS — A NEW CLASSIFICATION METHOD OF SOCIO-ECONOMIC PHENOMENA

Summary

Two manners of class separation in a unit set are known: 1) set division into homogeneous parts, and 2) grouping similar elements. Non-ihierarchical cluster analysis presented in the article is a classification method based on unit grouping.

It can be widely used and there is no need to accept preliminary assumptions on structure of examined set of elements. It is, however, in the case of hierarchical grouping techniques. The discussed mthod can be used not only as a method of hypotheses testing but above all as a methods of their creation.

Following steps exist when non-hierarchical cluster analysis is used: 1) determination of an approximate number, of classes (a simplified picture of distribution of points representing set elements in two-dimensional space is used to this end), 2) identification of cluster cores by: a) arbitrary choice of typical units from observations matrix, b) random choice of cores from values between the minimum and the maximum of each feature observed, 3) distribution of units to cores on the ground of calculated distance, and 4) statistical appraisal of solution correctness (f.ex. F Test).

Each step has been discussed in the first part of the article in detail.

An example of utilization of non-hierarchical cluster analysis for determination of types of socio-economic structure of the Wielkopolska villages was presented in the second part of the elaboration. The obtained results were compared with the country typology determined on the ground of the Czekanowski's diagram. It allowed to present advantages and imperfections of non-hierarchical cluster analysis.