Raport Badawczy

Research Report

RB/49/2009

Wybrane metody podziału i klasyfikacji w analizie struktur regionalnych

G. Petriczek

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Pracowni zgłaszający pracę:
dr inż. Jan W. Owsiński

Warszawa 2009

WYBRANE METODY PODZIAŁU I KLASYFIKACJI W ANALIZIE STRUKTUR REGIONALNYCH

GRAŻYNA PETRICZEK

Warszawa 2009

1. WSTĘP

M. Walesiak [1994], s. 53 zaproponował podział metod taksonomicznych (w wąskim znaczeniu) na trzy podstawowe grupy:

- metody hierarchiczne,
- obszarowe i gęstościowe,
- metody optymalizujące wstępny podział zbioru obiektów.

Wśród *metod hierarchicznych* wyróżnia się *metody aglomeracyjne* oraz *deglomeracyjne*. W praktycznych zastosowaniach przeważają metody aglomeracyjne, są też one najlepiej opracowane pod względem metodologicznym[52].

Metody aglomeracyjne zaczynają się zawsze od sytuacji, w której każdy obiekt badania O_i tworzy początkowo jedną klasę K_i, gdzie $i=1,2,...,n$. W związku z tym macierz e odległości przybierają postać:

$$[d_{ij}]=\begin{bmatrix} 0 & d(K_1,K_2) & \dots & d(K_1,K_n) \\ d(K_2,K_1) & 0 & \dots & d(K_2,K_n) \\ \dots & \dots & \dots & \dots \\ d(K_n,K_1) & d(K_n,K_2) & \dots & 0 \end{bmatrix}$$

Wszystkie hierarchiczne *metody klasyfikacji aglomeracyjnej* działają według centralnej procedury aglomeracyjnej. Algorytm tej procedury jest przedstawiony poniżej[53].

1. Tworzy się n-skupień, czyli każdy obiekt badania stanowi jedną klasę.
2. W macierzy odległości szuka się pary klas najbardziej podobnych (najmniej odległych od siebie). Załóżmy, że będą to klasy K_i oraz K_j.
3. Redukuje się liczbę klas o jeden, łącząc klasy K_i oraz K_j w nową klasę.
4. Przekształca się odległości stosownie do metody między połączonymi klasami K_i i K_j oraz pozostałymi klasami.
5. Powtarza się kroki 1-3 do chwili, gdy wszystkie obiekty znajdą się w jednej klasie.

Ogólna filozofia *metod obszarowych i gęstościowych* polega na tym, że wydzielonymi przy ich użyciu klasami są takie obszary w przestrzeni p-wymiarowej, które charakteryzują się większą gęstością obiektów i są oddzielone obszarami o mniejszej gęstości obiektów. Do metod tego typu należą m.in. *metoda kul*[59], *metoda taksonomii stochastycznej*[60], *metoda prostopadłościanów*[61] oraz *metoda grafowa Pluty*[62].

Punktem wyjścia *metod optymalizacji iteracyjnej* jest wstępny podział zbioru obiektów na l klas otrzymany przy użyciu dowolnej metody klasyfikacji lub ustalony losowo. Zadaniem tych metod jest poprawienie z punktu widzenia pewnej zdefiniowanej funkcji kryterium wstępnego podziału zbioru obiektów.

2. OMÓWIENIE WYBRANYCH METOD

2.1. METODA CZEKANOWSKIEGO

Przykład metod diagramowych stanowi metoda Czekanowskiego. Procedura zaproponowana przez J. Czekanowskiego (1913) jest najstarszą numeryczną procedurą taksonomiczną.
W metodach diagramowych stosuje się graficzną prezentację macierzy odległości zwaną diagramem.
Punkt wyjścia metody Czekanowskiego stanowi macierz odległości między obiektami $\boldsymbol{D}=[d_{ii'}]$, zdefiniowana za pomocą dowolnej metryki (Grabiński i in., 1989). Mierniki odległości w macierzy odległości $\boldsymbol{D}$ dzieli się na klasy podobieństwa obiektów[1]. Poszczególnym klasom podobieństwa obiektów przyporządkowuje się odpowiednie symbole graficzne, otrzymując nieuporządkowany diagram Czekanowskiego, co pozwala na wzrokową ocenę przebiegu porządkowania obiektów. Samo porządkowanie obiektów odbywa się poprzez porządkowanie diagramu. Polega ono na przestawianiu wierszy i odpowiadających im kolumn diagramu tak aby symbole graficzne reprezentujące możliwe najmniejsze odległości skupiały się wzdłuż głównej przekątnej, a w miarę oddalania się od głównej przekątnej pojawiały się symbole graficzne odpowiadające coraz większym odległościom. Kolejność uporządkowania obiektów jest określona przez kolejność odpowiadających im wierszy (kolumn).

Wyłącznie wzrokowa ocena poprawności uporządkowania diagramu ma charakter subiektywny i przy dużej liczbie obiektów może sprawiać trudności. Stąd też do oceny poprawności

porządkowania zaproponowane zostało pewne kryterium oparte na obiektywnej funkcji o postaci (Grabiński, 1977; Kolenda, 2006):

$$F^1 = \sum_{i=1}^{n}\sum_{i'>1}^{n} d_{ii'} w_{ii'} \rightarrow \max, \tag{2.1}$$

gdzie:

$w_{ii'}$ – wagi elementów macierzy odległości, zdefiniowane w oparciu o jeden z następujących wzorów:

$$w_{ii'} = \frac{|i-i'|}{n-1}, \tag{2.2}$$

$$w_{ii'} = \frac{1}{n(n-1)}\left[2n|i-i'-1| + i + i' - (i-i')^2\right], \tag{2.3}$$

$$w_{ii'} = \frac{1}{n(n-1)}\left[2n|i-i'| + 2 - i - i' - (i-i')^2\right]. \tag{2.4}$$

Wagi elementów macierzy odległości tworzą macierz wag o postaci:

$$\boldsymbol{W} = [w_{ii'}], \qquad i,i' = 1,2,\ldots,n. \tag{2.5}$$

Wagi w macierzy $\boldsymbol{W}$ są rozmieszczone zgodnie z pożądanym rozmieszczeniem elementów w macierzy odległości $\boldsymbol{D}$, czyli innymi słowy macierz $\boldsymbol{W}$ stanowi wzorzec dla docelowego uporządkowania diagramu powstałego z macierzy odległości $\boldsymbol{D}$. Porządkując diagram Czekanowskiego przestawimy w nim wiersze i odpowiednie kolumny w taki sposób aby były ułożone zgodnie ze wzorem wag w macierzy $\boldsymbol{W}$, co osiąga się maksymalizując funkcję poprawności uporządkowania (2.1).

2.2. METODY DENDRYTOWE

Metody dendrytowe opierają się na regułach i pojęciach teorii grafów. Przez graf G(O,Γ) będziemy rozumieli zbiór porządkowanych obiektów, a przez Γ relację przyporządkowującą każdemu obiektowi ze zbioru O obiekt najbliżej od niego położony (Grabiński i in., 89,). W graficznej prezentacji grafu poszczególnym obiektom odpowiadają punkty nazywane wierzchołkami, a odcinki łączące te punkty, o długościach równym odległościom pomiędzy obiektami, nazywane są łukami lub wiązadłami.

Dendryt jest grafem spójnym i otwartym. Graf nazywany jest spójnym, jeżeli jego każda para wierzchołków jest połączona nieprzerwanym ciągiem wiązadeł. Natomiast graf otwarty charakteryzuje się brakiem cykli i pętli. Cyklem jest skończony ciąg połączonych ze sobą wiązadeł, w którym początkowy wierzchołek pierwszej krawędzi stanowi jednocześnie końcowy wierzchołek ostatniej krawędzi. Pętlą natomiast nazywamy cykl składający się tylko z jednego wiązadła.

Uporządkowanie dendrytowe polega na przyporządkowaniu obiektom poszczególnych wierzchołków dendrytu. Przykłady metod dendrytowych stanowią **taksonomia wrocławska** oraz metoda Prima.

2.3. TAKSONOMIA WROCŁAWSKA

Zasady metody taksonomii wrocławskiej zostały opracowane przez członków Grupy Zastosowań Państwowego Instytutu Matematycznego we Wrocławiu (Florek i in., 1951). Sam proces budowy dendrytu wrocławskiego, w ramach powyższej metody, jest procesem wieloetapowym.

- W pierwszym etapie szukamy dla każdego obiektu O_i obiektu $O_{i'}$ najbardziej do niego podobnego. W tym celu w każdym wierszu (kolumnie) macierzy odległości $\boldsymbol{D}$ wyznaczamy najmniejszy element:

$$d_{ii'} = \min_{i'}\{d_{ii'}\}, \qquad i,i'=1,2,\ldots,n;\ i\neq i'. \tag{2.6}$$

- Otrzymane pary najbardziej podobnych do siebie obiektów przedstawiamy w postaci grafu niezorientowanego, tzn. grafu, w którym wierzchołki odpowiadające tym obiektom są połączone wiązadłami bez zaznaczania kierunku połączenia. Długości tych krawędzi są proporcjonalne do odległości między obiektami. Wśród wyznaczonych par połączeń mogą znajdować się połączenia występujące dwukrotnie.

- Ponieważ kolejność połączeń w dendrycie nie odgrywa roli jedno z podwójnych połączeń jest eliminowane. Ponadto w łączeniu mogą występować wielokrotnie te same obiekty, a w dendrycie dany obiekt może występować tylko jeden raz. Dla zapewnie-

nia powyższego warunku połączenia te łączone są wtedy w zespoły, nazywane skupieniami.

- Po utworzeniu w powyższy sposób grafu sprawdzamy czy jest on spójny.
  - jeżeli uzyskaliśmy spójny graf budowa dendrytu została zakończona.
  - jeżeli natomiast otrzymany graf nie jest spójny to jego poszczególne składowe (skupienia) łączy się w większe zespoły.

- Poszczególne skupienia łączymy ze sobą w miejscach określonych przez minimalną odległość między nimi. Tworzymy w ten sposób skupienia 2-giego rzędu.
  - znajdujemy w tym celu najmniejszą odległość każdego obiektu jednego skupienia od obiektów należących do pozostałych skupień.
  - z odległości tych wybieramy odległość najmniejszą, która zostaje wiązadłem łączącym skupienia.

- Jeżeli graf w dalszym ciągu nie jest spójny proces ten jest kontynuowany poprzez tworzenie skupień wyższego rzędu.
- Otrzymanie spójnego grafu kończy proces tworzenia dendrytu.

Uzyskane w ten sposób uporządkowanie dendrytowe jest najkrótsze (suma długości wiązadeł dendrytu jest najmniejsza) ze wszystkich możliwych uporządkowań dendrytowych.

2.4 METODY AGLOMERACYJNE

Metody aglomeracyjne prowadzą do utworzenia drzewka połączeń (tzw. dendrogramu), które stanowi ilustrację graficzną sposobu i hierarchii łączenia obiektów, ze względu na zmniejszające się podobieństwo między obiektami włączonymi do drzewka w kolejnych etapach, a obiektami wcześniej włączonymi do drzewka. Hierarchia tych połączeń pozwala na określenie wzajemnego położenia względem siebie obiektów oraz grup obiektów powstających na kolejnych etapach tworzenia drzewka (Lance i Williams, 1967 i 1968; Sneath i Sokal, 1973; Jajuga, 1990). Grupy podobnych do siebie obiektów tworzą na tym hierarchicznym drzewku oddzielne gałęzie.

Punktem wyjścia metod aglomeracyjnych jest założenie, że każdy obiekt stanowi odrębną, jednoelementową grupę (G_r, $r=1,2,...,z$). Następnie, w kolejnych krokach, łączymy ze sobą grupy obiektów najbardziej do siebie podobnych ze względu na wartości opisujących je zmiennych. Miarą tego podobieństwa są odległości między grupami obiektów.

W pierwszym kroku odległości między jednoelementowymi grupami obiektów $G_1,...,G_n$ są elementami wejściowej macierzy odległości $\boldsymbol{D}$. Obiekty traktujemy tutaj w sposób węższy jako obiekty przestrzenne. Tym samym w macierzy $\boldsymbol{D}$ szukamy najmniejszej odległości pomiędzy tymi grupami obiektów:

$$d_{rr'} = \min_{ii'}\{d_{ii'}\}, \qquad i=1,2,...,n_r;\ i'=1,2,...,n_{r'};\ r,r'=1,2,...,n;\ r\neq r'. \tag{2.7}$$

gdzie:

$d_{rr'}$ – odległość r-tej od r'-tej grupy.

Obiekty najbardziej do siebie podobne łączymy w jedną grupę, co powoduje zmniejszenie wyjściowej liczby grup o jeden, rozpoczynając budowę drzewka połączeń. Następnie wyznaczamy odległości nowo utworzonej grupy obiektów od wszystkich pozostałych grup obiektów. Odległości te wstawia się do macierzy odległości $\boldsymbol{D}$ w miejsce wierszy i kolumn odpowiadających obiektom (grupom obiektów) połączonym w jedną grupę. Procedurę łączenia grup obiektów powtarza się do momentu gdy tworzą one jedną grupę (zostało utworzone pełne drzewko połączeń), czyli n-1 razy. Po każdym etapie grupowania określamy odległości nowopowstałej grupy obiektów od pozostałych grup obiektów. Odległości te tworzą nową, aktualną na danym etapie grupowania, macierz odległości o coraz mniejszym wymiarze (n-u) (n-u), gdzie u jest u-tym etapem łączenia grup obiektów.

Ogólna formuła wyznaczania odległości nowo powstałej grupy obiektów $G_{r''}$, poprzez połączenie grup obiektów G_r i $G_{r'}$, od pozostałych grup obiektów $G_{r'''}$, przy tworzeniu drzewka połączeń ma następującą postać (Lance i Williams, 1967 i 1968):

$$d_{r'''r''} = \alpha_r d_{r'''r} + \alpha_{r'} d_{r'''r'} + \beta d_{rr'} + \gamma\left|d_{r'''r} - d_{r'''r'}\right|, \tag{2.8}$$

gdzie:

$\alpha_r, \alpha_{r'}, \beta, \gamma$ - współczynniki przekształceń odmienne dla różnych metod drzewkowych.

Poszczególne metody drzewkowe różnią się między sobą sposobami wyznaczania odległości między obiektami (Wishart, 1969).

Do najczęściej stosowanych w praktyce metod aglomeracyjnych należą:

- metoda najbliższego sąsiedztwa (metoda pojedynczego wiązania),
- metoda najdalszego sąsiedztwa (metoda pełnego wiązania),
- metoda średniej międzygrupowej (metoda średnich połączeń),
- metoda mediany,
- metoda środka ciężkości,
- metoda Warda.

2.4.1 Metoda najbliższego sąsiedztwa

W metodzie tej odległość między dwoma grupami obiektów jest definiowana jako odległość pomiędzy najbliższymi obiektami (najbliższymi sąsiadami) należącymi do dwóch różnych grup obiektów (Sneath i Sokal, 1973). Innymi słowy jest to najmniejsza z odległości pomiędzy dwoma dowolnymi obiektami należącymi do poszczególnych grup (rys. 2.7), co od strony formalnej możemy przedstawić następująco:

$$d_{rr'} = \min_{ii'}\{d_{ii'}(\mathbf{O}_i \in G_r, \mathbf{O}_{i'} \in G_{r'})\}, \qquad i=1,2,...,n_r;\ i'=1,2,...,n_{r'};\ r,r'=1,2,...,z;\ r\neq r', \quad (2.9a)$$

gdzie:

$$\mathbf{O}_i=[z_{ij}], \qquad j=1,2,...m. \qquad (2.9b)$$

Parametry przekształceń mają wartości: α_r=0,5, $\alpha_{r'}$=0,5, β=0 i γ=0,5. W wyniku stosowania tej metody obiekty łączą się w grupy tworzące "łańcuchy

2.4.2. Metoda najdalszego sąsiedztwa

Odległość między dwoma grupami obiektów jest tutaj określana jako odległość pomiędzy najdalszymi obiektami (najdalszymi sąsiadami) należącymi do różnych grup obiektów. Ozna-

cza to, że jest to największa spośród odległości pomiędzy dwoma dowolnymi obiektami należącymi do różnych grup:

$$d_{rr'} = \max_{ii'}\{d_{ii'}(\mathbf{O}_i \in G_r, \mathbf{O}_{i'} \in G_{r'})\},\ i=1,2,...,n_r;\ i'=1,2,...,n_{r'};\ r,r'=1,2,...,z;\ r\neq r'. \qquad (2.10)$$

Parametry przekształceń przyjmują wartości: α_r=0,5, $\alpha_{r'}$=0,5, β=0 i γ=-0,5. Metoda ta prowadzi do łączenia się obiektów w grupy tworzące "kępki".

2.4.3. Metoda średniej międzygrupowej

W metodzie tej odległość między dwoma grupami obiektów obliczana jest jako średnia arytmetyczna odległości między wszystkimi parami obiektów należących do dwóch różnych grup:

$$d_{rr'} = \frac{1}{n_r n_{r'}} \sum_{i'=1}^{n_{r'}} \sum_{i=1}^{n_r} d_{ii'}(\mathbf{O}_i \in G_r, \mathbf{O}_{i'} \in G_{r'}), \qquad r,r'=1,2,...,z;\ r\neq r'. \qquad (2.11)$$

Parametry przekształceń przyjmują wartości: $\alpha_r = \frac{n_r}{n_r + n_{r'}}$, $\alpha_{r'} = \frac{n_{r'}}{n_r + n_{r'}}$, β=0 i γ=0.

Powyższa metoda może dawać w efekcie drzewka połączeń składające się zarówno z grup obiektów tworzących "kępki" jak i z grup obiektów tworzących "łańcuchy".

2.4.4. Metoda środków ciężkości

Odległość między dwoma grupami w tej metodzie określona jest jako odległość między środkami ciężkości tych grup:

$$d_{rr'} = d_{i^c i'^c}\left(\mathbf{O}_{i^c} = \overline{\mathbf{O}}_r \in G_r, \overline{\mathbf{O}}_{i'^c} = \mathbf{O}_{r'} \in G_{r'}\right), i=1,2,...,n_r; i'=1,2,...,n_{r'};\ r,r'=1,2,...,z; \qquad (2.12)$$

gdzie:

$d_{i^c i'^c}$ - odległość środka ciężkości r-tej grupy od środka ciężkości r'-tej grupy.

$\overline{\mathbf{O}}_{i^c}, \overline{\mathbf{O}}_{i'^c}$ – środki ciężkości odpowiednio r-tej i r'-tej grupy obiektów, przy czym:

$$\mathbf{O}_{i^{\prime c}} = \overline{\mathbf{O}}_r = \frac{1}{n_r}\sum_{i=1}^{n_r}\mathbf{O}_i\,, \qquad \mathbf{O}_{i^{\prime\prime c}} = \overline{\mathbf{O}}_{r^{\prime}} = \frac{1}{n_{r^{\prime}}}\sum_{i^{\prime}=1}^{n_{r^{\prime}}}\mathbf{O}_{i^{\prime}}\,.$$

Parametry przekształceń przyjmują wartości:

$$\alpha_r = \frac{n_r}{n_r + n_{r^{\prime}}},\ \alpha_{r^{\prime}} = \frac{n_{r^{\prime}}}{n_r + n_{r^{\prime}}},\ \beta = \frac{-n_r n_{r^{\prime}}}{(n_r + n_{r^{\prime}})^2} \text{ i } \gamma = 0.$$

2.5. METODA WARDA

W metodzie tej odległości między dwoma grupami obiektów nie można przedstawić wprost za pomocą odległości pomiędzy obiektami należącymi do tych grup. Dwie grupy obiektów przy tworzeniu drzewka połączeń, na dowolnym etapie, są łączone w jedną grupę tak, aby zminimalizować sumę kwadratów odchyleń wszystkich obiektów z tych dwóch grup od środka ciężkości nowej grupy, która powstanie w wyniku połączeń tych dwóch grup. Oznacza to, że na każdym etapie łączenia grup obiektów, ze wszystkich możliwych do łączenia grup obiektów, łączy się w jedną grupę te grupy, które w rezultacie tworzą grupę obiektów o najmniejszym zróżnicowaniu ze względu na opisujące je zmienne. Miarą tego zróżnicowania jest kryterium *ESS* (*Error Sum of Squares*) sformułowane przez J. H. Warda (1963) o postaci:

$$\mathrm{ESS} = \sum_{i^{\prime\prime}=1}^{n_{r^{\prime\prime}}} d^2_{i^{\prime\prime}i^{\prime\prime c}}\left(\mathbf{O}_{i^{\prime\prime}} \in G_{r^{\prime\prime}}, \mathbf{O}_{i^{\prime\prime c}} = \overline{\mathbf{O}}_{r^{\prime\prime}} \in G_{r^{\prime\prime}}\right), \tag{2.13}$$

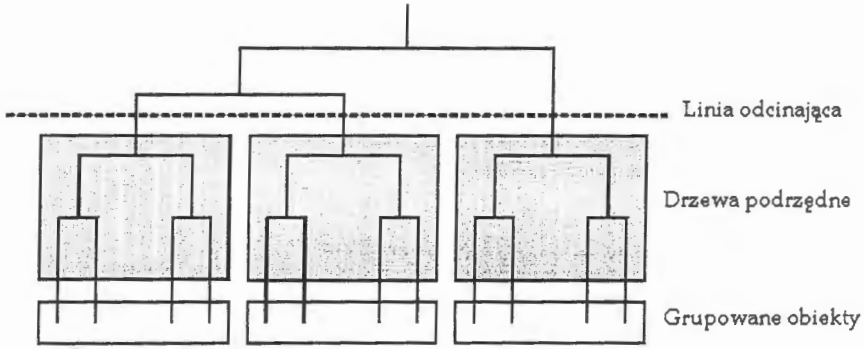
gdzie:

$d_{i^{\prime\prime}i^{\prime\prime c}}$ – odległość $i^{\prime\prime}$-tego obiektu należącego do nowopowstałej $r^{\prime\prime}$-tej grupy od środka ciężkości tej grupy.

$$\mathbf{O}_{i^{\prime\prime c}} = \overline{\mathbf{O}}_{r^{\prime\prime}} = \frac{1}{n_{r^{\prime\prime}}}\sum_{i^{\prime\prime}=1}^{n_{r^{\prime\prime}}}\mathbf{O}_{i^{\prime\prime}}\,. \tag{2.14}$$

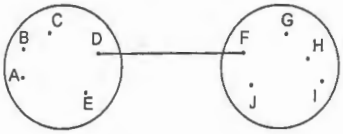
W metodzie Warda parametry przekształceń przyjmują postać:

$$\alpha_r = \frac{n_r + n_{r^{\prime\prime\prime}}}{n_r + n_{r^{\prime}} + n_{r^{\prime\prime\prime}}},\ \alpha_{r^{\prime}} = \frac{n_{r^{\prime}} + n_{r^{\prime\prime\prime}}}{n_r + n_{r^{\prime}} + n_{r^{\prime\prime\prime}}},\ \beta = \frac{-n_{r^{\prime\prime\prime}}}{n_r + n_{r^{\prime}} + n_{r^{\prime\prime\prime}}} \text{ i } \gamma = 0.$$

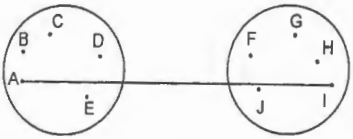


Przykładowy dendrogram

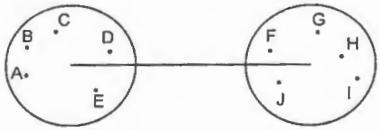
Odległości międzygrupowe w wybranych metodach aglomeracyjnych



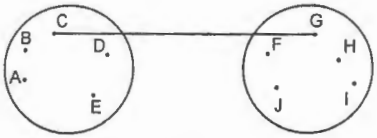
Metoda najbliższego sąsiedztwa



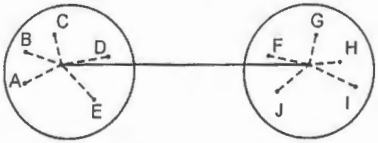
Metoda najdalszego sąsiedztwa



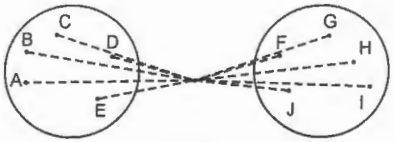
Metoda średnich połączeń



Metoda mediany



Metoda środków ciężkości



Metoda Warda

2.6. METODY OPTYMALIZACJI DANEGO GRUPOWANIA OBIEKTÓW

Ogólna charakterystyka

- punktem wyjścia metod optymalizacyjnych jest ustalenie pożądanej liczby grup obiektów, które chcemy utworzyć
- ustalamy wstępny skład poszczególnych grup:
  - w sposób losowy
  - korzystając z ocen ekspertów
  - poprzez wykorzystanie arbitralnie wybranej zmiennej
  - przyjmując jako wstępne grupowanie, grupowanie otrzymane za pomocą dowolnej metody taksonomicznej
  - porządkując obiekty według ich odległości od środka ciężkości poszczególnych grup obiektów. Środkami ciężkości grup obiektów stają się obiekty o numerach określonych za pomocą wzoru: $1+(r-1)\left(\frac{n}{z}\right)$, gdzie r jest kolejnym numerem grupy
- następnie poprawiamy dobroć wstępnego grupowania obiektów poprzez optymalizację grupowania polegającą na przesuwaniu obiektów między grupami

2.6.1. METODA k-ŚREDNICH

Metoda k – średnich należy do metod optymalizacyjno-iteracyjnych. Istota tej grupy metod polega na tym, że optymalizowana jest pewna funkcja jakości podziału obiektów.

Metody podziałowe polegają na dzieleniu całego zbioru obiektów zgodnie z ogólną zasadą maksymalizacji wariancji pomiędzy poszczególnymi grupami, przy jednoczesnej minimalizacji wariancji wewnątrz badanych grup. Idea metody k-średnich została opracowana w latach pięćdziesiątych przez T. Daleniusa, [Cox, 1957] dla jednowymiarowych zmiennych

Uogólnienie dla przypadku wielowymiarowego przedstawił G. S. Sebestyen [Sebestyen, 1962]. Autorstwo metody k -średnich, przypisuje się jednak J. McQueen'owi [McQueen, 1967], który rozpatrywał efektywność tejże metody z punktu widzenia losowego doboru obiektów do wyróżnionych grup. [Grabiński, 1992].

Ogólny schemat postępowania jest następujący:

- na wstępie ustalamy podział obiektów na grupy oraz liczbę iteracji, w których dążymy do optymalizacji grupowania
- następnie obliczamy wartość funkcji kryterium dobroci grupowania, którą stanowi np. stosunek zróżnicowania międzygrupowego do zróżnicowania wewnątrzgrupowego
  - miara zróżnicowania międzygrupowego najczęściej jest definiowana jako suma odległości środków ciężkości grup obiektów od środka ciężkości wszystkich badanych obiektów
  - ocenę zróżnicowania wewnątrzgrupowego stanowi suma odległości wewnątrzgrupowych obiektów od środków ciężkości grup, do którego zostały one sklasyfikowane. Wartość funkcji kryterium może mieć także postać statystyki F stosowanej w analizie wariancji
- w kolejnym kroku obliczamy środki ciężkości dla poszczególnych grup i klasyfikujemy obiekty do grup na podstawie minimalizacji ich odległości od środków grup
- następnie sprawdzamy czy wartość funkcji kryterium nie zwiększyła się
  - gdy zmiana taka nie nastąpiła kończymy procedurę przyjmując, że dane grupowanie jest optymalne
  - w sytuacji przeciwnej przechodzimy do kolejnej iteracji, sprawdzając czy przesunięcia obiektów między grupami nie powodują wzrostu wartości funkcji kryterium dobroci grupowania
  - procedurę kontynuujemy do momentu gdy wartość funkcji kryterium dobroci grupowania nie zwiększa się albo gdy osiągnęliśmy założoną liczbę iteracji

Ogólna idea tych procedur polega na poprawianiu danego podziału obiektów z punktu widzenia odpowiednio zdefiniowanego kryterium optymalności podziału.

Zakładamy, $k \in (2, n-1)$, gdzie n jest liczbą obiektów.

Wariant metody k-średnich można opisać następująco.

Niech **X1, X2, X3 , ..., Xn** będą obiektami m cechowymi. (to znaczy **X**1 = [x11,..., x1m]).

- Na początku ustala się wyjściową macierz środków ciężkości grup

$B = [\bar{x}_{ij}]$ (1 = 1, ..., p: j = 1,..... m)

gdzie m – liczba zmiennych

- Dla każdej z grup obliczamy średnią (położenie centroidu). Wyznacza się odległości pierwszej nieprzydzielonej jednostki od środków ciężkości poszczególnych grup i kwalifikuje ją do grupy najbliżej położonej.

Następnie wyznacza się wartość wyjściowego błędu podziału obiektów między k grup

$$e = \sum_{i=1}^{n} d_{il}^2$$

gdzie:

d^2_{il} odległość Euklidesa między i-tym obiektem a najbliższym l-tym środkiem ciężkości:

$$d_{il}^2 = \sum_{j=1}^{m} (x_{ij} - \bar{x}_{lj})^2 \quad (i = 1,\ldots,n) \tag{2.15}$$

Zestaw odległości euklidesowych obliczany jest pomiędzy poszczególnymi elementami zbioru a kolejnymi centroidami.

Dla pierwszego obiektu określa się zmiany błędu podziału wynikające z przyporządkowania go kolejno do wszystkich aktualnie występujących grup:

$$\Delta e_l^{(1)} = \frac{n_k d_{1k}^2}{n_k + 1} - \frac{n_{k_1} d_{1k_1}^2}{n_{k_1} - 1} \tag{2.16}$$

gdzie:

n_k – liczebność k - tej grupy,

d_{1k} – odległość pierwszego obiektu od środka ciężkości k - tej grupy,

n_{kl} - liczebność grupy zawierającej pierwszy obiekt,

d_{1kl} – odległość pierwszego obiektu od najbliższego środka ciężkości.

- Jeżeli minimalna wartość wyrażenia Δe dla wszystkich l jest ujemna, to pierwszy obiekt przypisuje się do grupy, dla której Δe = min.
- Następnie powtarza się obliczenia to znaczy od nowa oblicza się środki ciężkości grup **B** uwzględniając dokonaną transformację obiektu oraz wyznacza aktualną wartość błędu podziału.

- Jeżeli minimalna wartość wyżej przedstawionego wyrażenia jest dodatnia lub równa zero, to nie dokonujemy już żadnych zmian.
- Operacje opisane powyżej powtarza się dla każdego następnego obiektu.
- Gdy nie obserwujemy już żadnych przesunięć obiektów z grupy do grupy, czyli gdy każdy element jest w grupie, w której centroid jest mu najbliższy, wówczas postępowanie się kończy w pierwszej wersji podziału.
- W przeciwnym wypadku rozpoczyna się następną iterację, aż do momentu, w którym ich liczba nie przekroczy zadanej wartości [Zeliaś i in., 1989, Witkowska, 2002].

2.6.2. MODYFIKACJA METODY K - ŚREDNICH.

Obserwowane obiekty **X**1, **X**2,...,**X**n są obiektami m cechowymi, to znaczy **X**1 = (Xi1,...,Xi$_m$), gdzie i = 1,...,N. Ponieważ liczba k skupień jest z góry ustalona, szukamy najlepszego podziału **J**(k) = {G1, G2,..., Gk}, którym będzie podział zbioru {1,..., N} na k rozłącznych podzbiorów. Wybieramy najlepszy spośród wszystkich uzyskanych podziałów, to znaczy taki, dla którego zróżnicowanie wewnątrzgrupowe było najmniejsze oraz zmienność pomiędzy grupami była jak największa czyli oznaczając taki podział przez **J***(k) (to taki podział na k grup, że zróżnicowanie międzygrupowe w stosunku do zróżnicowania wewnątrzgrupowego jest największe.)

Jako miernik zróżnicowania międzygrupowego przyjmuje się:

$$S^2_{A(J(k))} = \frac{1}{k-1}\sum_{i=1}^{k}\left\|\bar{x}_i - \bar{x}_{J(k)}\right\|^2 = \frac{1}{k-1}\sum_{i=1}^{k} d_i^2 \qquad (2.17)$$

$\bar{x}_{J(k)} = \frac{1}{k}\sum_{i=1}^{k}\bar{x}_{G_i}$ środek ciężkości proponowanego podziału J(k)

$\bar{x}_{G_i}$ - środek ciężkości i-tej grupy.

Jako miernik zróżnicowania wewnątrzgrupowego:

$$S^2_{E(J(k))} = \frac{1}{N-k-1}\sum_{i=1}^{k}\left(\sum_{j=1}^{k_i}\left\|x_{ij} - \bar{x}_{G_i}\right\|^2\right) = \frac{1}{N-k-1}\sum_{i=1}^{k}\left(\sum_{j=1}^{k_i} d_{ij}^2\right) \quad (2.18)$$

$$f(k) = \frac{S^2_{A(J^*(k))}}{S^2_{E(J^*(k))}}$$

Funkcja kryterium to ogólna suma odległości wewnątrzgrupowych liczonych od środka grup, których współrzędne wyznaczono jako średnie arytmetyczne wartości cech obiektów należących do danej podgrupy. Jako optymalny podział J*(k) obiektów na skupienia wybiera się ten, dla którego funkcja osiąga maksimum [Pietrzykowski i in., 2005].

3. PRZYKŁADY WYKORZYSTANIA WYBRANYCH METOD TAKSONOMICZNYCH W ANALIZIE STRUKTUR REGIONALNYCH

3.1. METODA CZEKANOWSKIEGO

W artykule J. Górka i M. Chmurska przedstawiły zastosowanie **metody Czekanowskiego** do oceny i zakwalifikowania badanych powiatów województwa wielkopolskiego do grup typologicznych, w zależności od poziomu rozwoju turystyczno-rekreacyjnego.

Badania z dziedziny turystyczno-rekreacyjnej dotyczą regionów posiadających walory turystyczne. Stąd też podjęcie badań związane jest z przeprowadzeniem syntetycznej analizy badanych jednostek. Przeprowadzenie takiej analizy wspomnianego zjawiska może komplikować fakt dużej różnorodności opisywanych problemów turystyczno-rekreacyjnych.

Syntetyczna ocena jednostek turystyczno-rekreacyjnych w ujęciu przekrojowym wymaga zastosowania odpowiednich metod badawczych, które mogłyby być wykorzystane w warunkach złożoności analizowanych zjawisk. Bardzo cenne są analizy porównawcze dokonywane na podstawie metod statystyczno-matematycznych. Porównywanie bowiem stanowi podstawę oceny stanu jednostek turystyczno-rekreacyjnych, jak i relacji między tymi jednostkami [Nowak K. 1990]. Najczęściej jest to porównywanie zbioru cech analizowanego systemu między sobą lub jest to określenie relacji między cechami a stanem tej samej jednostki z przeszłości.

Porównawcze badania turystyczno-rekreacyjne dotyczą głównie jednostek administracyjnych, takich jak: gmina, powiat, województwo, kraj. Uzyskane wyniki badań mogą stanowić bardzo cenne źródło informacji dla terenowych organów administracji państwowej, gdyż mogą być wykorzystane jako podstawa racjonalnego planowania rozwoju turystyczno-rekreacyjnego w skali całego kraju lub wydzielonego obszaru.

Metodę tę używamy do łączenia jednostek terytorialnych w jednorodne rejony. Zaletą diagramu jest to, że uwypukla on najważniejsze związki i podobieństwa badanych obiektów, a równocześnie ujmuje wszystkie szczegółowe powiązania między jednostkami obszarowymi

Punktem wyjścia dla sporządzenia diagramu Czekanowskiego jest macierz odległości Euklidesowych między klasyfikowanymi obiektami. Odległości te są dzielone na klasy, które stanowią przedziały podobieństwa obiektów. Po ustaleniu skali podobieństwa, poszczególnym klasom przyporządkowuje się odpowiednie symbole graficzne odpowiadające poziomowi odległości między obiektami. Następnie macierz tę przekształca się w ten sposób, że poszczególne odległości zastępuje się symbolami, w wyniku czego otrzymujemy nieuporządkowany diagram Czekanowskiego. Wyjściowy diagram należy przekształcić tak, aby znaki graficzne oznaczające najmniejsze odległości skupiały się wzdłuż głównej przekątnej, wyznaczając w ten sposób grupę typologiczną, obejmującą jednostki najmniej zróżnicowane co do wartości opisujących je cech. Polega to na jednoczesnym przestawieniu wierszy i kolumn odpowiadających klasyfikowanym obiektom. W wyniku takich posunięć uzyskujemy uporządkowany diagram Czekanowskiego.

Głównym przeznaczeniem metody Czekanowskiego jest wyodrębnienie podzbiorów jednorodnych obiektów. Na grupy obiektów podobnych wskazują w uporządkowanym diagramie Czekanowskiego zespoły symboli obrazujących najmniejszą odległość między obiektami. Każdy zespół takich symboli, skupionych wzdłuż głównej przekątnej diagramu, wyznacza grupę typologiczną, obejmującą jednostki najmniej zróżnicowane co do wartości opisujących je cech. Ogromną zaletą metody Czekanowskiego jest fakt, że podczas klasyfikacji rozpatruje się całą macierz odległości.

Na podanym przykładzie opisu 31 powiatów województwa wielkopolskiego zaprezentowano metodę Czekanowskiego. Powiatom badanego województwa przypisano odpowiednie numery.

Każdy powiat został opisany takimi samymi 10 cechami diagnostycznymi **X1-X10**:

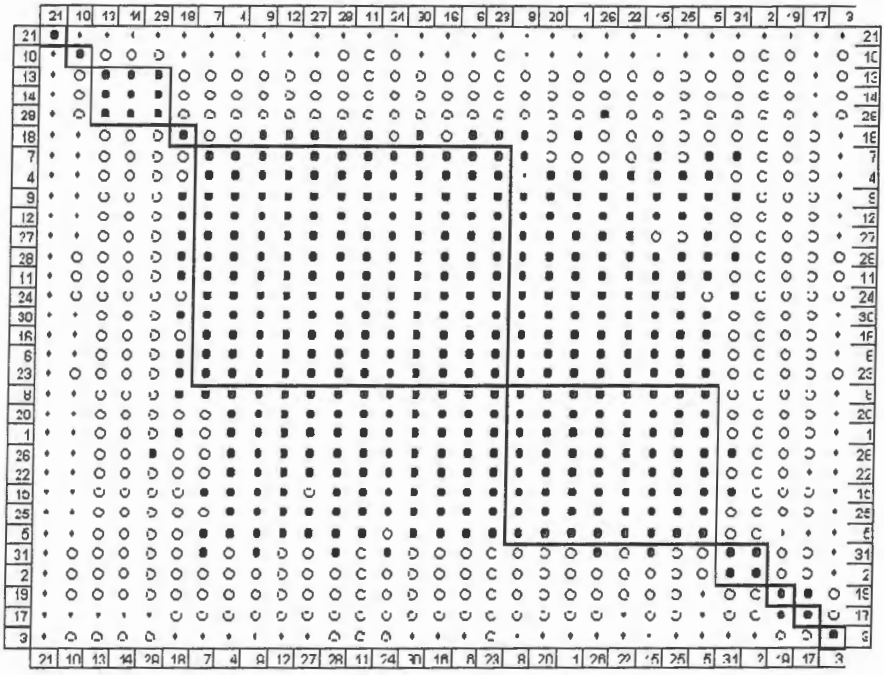
X1 Powierzchnia (w km^2)
X2 Ludność (ogółem)
X3 Drogi powiatowe (ogółem)
X4 Powierzchnia obszarów prawnie chronionych (w km^2)
X5 Obiekty noclegowe turystyki
X6 Dochody budżetów powiatów (w tys. zł)
X7 Nakłady inwestycyjne na ochronę środowiska (w tys. zł)
X8 Miejsca noclegowe turystyki
X9 Udzielone noclegi
X10 Tereny rekreacji i wypoczynku (w km^2)

Wartości cech sprowadzono do porównywalności za pomocą standaryzacji, następnie na podstawie tych danych obliczono macierz odległości, a kolejnym etapem było ustalenie klas. Określono trzy klasy podobieństwa badanych obiektów:

Klasy/ Classes	Przedziały/Ranges			Kod/ Code
1	0,00	-	3,00	●
2	3,01	-	6,00	○
3	6,01	-	powyżej/ and more	•

Uporządkowane diagramy Czekanowskiego pozwolą wyodrębnić obiekty o podobnej strukturze. Zbudowano diagramy dla badanych lat 1999-2001. Przedstawiony niżej diagram dotyczy 1999 roku.

W przedstawionym przykładzie uporządkowane diagramy Czekanowskiego pozwolą wyodrębnić 6 grup typologicznych, złożonych z powiatów o podobnej strukturze cech turystyczno-rekreacyjnych.



Rysunek 1. Uporządkowany diagram Czekanowskiego 1999 rok

W latach 1999-2001 rozkład powiatów w poszczególnych grupach kształtował się następująco:

- W roku 1999 do pierwszej (A_1) najsilniejszej grupy należały powiaty: gostyński, jarociński, kaliski, kolski, kościański, krotoszyński, obornicki, ostrzeszowski, słupecki, szamotulski, turecki, wągrowiecki, wrzesiński. Drugą (A_2) grupę, silnie nawiązującą do pierwszej, tworzyły powiaty: chodzieski, grodziski, kępiński, nowotomyski, pleszewski, rawicki, średzki, śremski. Do trzeciej (A_3) należały powiaty: koniński, leszczyński, międzychodzki oraz wolsztyński. Kolejną czwartą (A_4) grupę stanowiły powiaty: złotowski i czarnkowsko-trzcianecki. Natomiast piątą (A_5) powiaty: gnieźnieński, ostrowski i pilski. Ostatnią szóstą (A_6) grupę, najmniej liczną, tworzył powiat poznański.

- W roku 2000 do pierwszej (A_1) najsilniejszej grupy należały powiaty: chodzieski, gostyński, jarociński, kolski, kościański, krotoszyński, nowotomyski, obornicki, pleszewski, rawicki, szamotulski, średzki, śremski, wągrowiecki, wrzesiński. Drugą (A_2) grupę, silnie nawiązującą do pierwszej podobnie jak w roku poprzednim, tworzyły powiaty: grodziski, kaliski, kępiński, ostrzeszowski, słupecki, złotowski. Do trzeciej (A_3) należały powiaty: gnieźnieński, koniński, leszczyński, międzychodzki oraz wolsztyński. Kolejną czwartą (A_4) grupę stanowiły powiaty: czarnkowsko-trzcianecki, ostrowski i pilski. Do ostatnich dwóch grup – piątej (A_5) i szóstej (A_6) – należały kolejno powiaty: turecki i poznański.

- W roku 2001 do pierwszej (A_1) najsilniejszej grupy należały powiaty: gostyński, grodziski, jarociński, kolski, krotoszyński, nowotomyski, obornicki, pleszewski, rawicki, średzki, śremski, turecki, wągrowiecki, wrzesiński. Drugą (A_2) grupę, silnie nawiązującą do pierwszej, tworzyły powiaty: chodzieski, kaliski, kępiński, ostrzeszowski, słupecki, szamotulski. Do trzeciej (A_3) należały powiaty: koniński, leszczyński, międzychodzki oraz wolsztyński. Kolejną czwartą (A_4) grupę stanowiły powiaty: czarnkowsko-trzcianecki i złotowski. Natomiast piątą (A_5) powiaty: gnieźnieński, kościański, ostrowski i pilski. Ostatnią szóstą (A_6) grupę, najmniej liczną, tworzył powiat poznański.

3.2. TAKSONOMIA WROCŁAWSKA

W pracy () Elżbieta Badach podała przykład zastosowania **taksonomii wrocławskiej** w badaniach populacji osób bezrobotnych w województwie małopolskim składającym się z 22 powiatów.

Materiał źródłowy wykorzystany do badań stanowiły dane gromadzone i udostępnione przez powiatowe urzędy pracy działające na terenie województwa małopolskiego. Wytypowano wstępnie 6 zmiennych, które charakteryzują zbiorowość bezrobotnych zarejestrowanych w danym urzędzie. Są to następujące wskaźniki struktury:

X1 - odsetek bezrobotnych nie przekraczających 25 roku życia,
X2 - odsetek bezrobotnych po 50 roku życia,
X3 - odsetek osób bezrobotnych do 27 roku życia z wyższym wykształceniem,
X4 - odsetek bezrobotnych bez kwalifikacji zawodowych,
X5 - odsetek bezrobotnych samotnie wychowujących dziecko do lat 7,
X6 - udział osób niepełnosprawnych w grupie zarejestrowanych bezrobotnych
(rezultacie 5 zmiennych)

Wszystkie wyróżnione kategorie osób bezrobotnych zostały uznane w myśl ustawy *o promocji zatrudnienia i instytucjach rynku pracy* (z 1 czerwca 2004 roku Dz.U) za grupy bezrobotnych szczególnie zagrożonych na rynku pracy, do których kierowane są specjalne formy pomocy.

Obiekty (czyli powiaty, utożsamiane w tym przypadku z lokalnymi rynkami pracy województwa małopolskiego) opisane przy pomocy zmiennych, poddano analizie. Jej celem było wyodrębnienie grup obiektów podobnych pod względem zespołu rozpatrywanych cech populacji bezrobotnych.

Posłużono się w tym celu **metodą taksonomii wrocławskiej,** pochodzącą z grupy taksonomicznych metod dendrytowych, opierających się na pojęciach z zakresu teorii grafów. Zastosowanie tej metody wymaga wstępnych przekształceń (standaryzacji) danych wejściowych zebranych w macierzy wymiaru $N \times L$,
gdzie: N - liczba obiektów poddanych analizie,
L -. liczba zmiennych uwzględnionych w badaniu.
Standaryzacji dokonano według wzoru:

$$z(i) = \frac{x(i) - \overline{x}}{S}$$

gdzie:
z(i) . zmienna po standaryzacji,
x(i) . realizacja i-tej cechy,
$\overline{x}$ - średnia wartość cechy w analizowanej próbie,
S - odchylenie standardowe z próby

Procedura taka umożliwia przedstawienie rozpatrywanych zmiennych w jednolitej skali. W przypadku zaniechania tej czynności na analizę miałyby decydujący wpływ zmienne o najwyższym zakresie wartości.
Dla przekształconych zmiennych oblicza się następnie macierz odległości (w tym przypadku euklidesowych) między obiektami, która stanowi punkt wyjścia do budowy dendrytu wrocławskiego.
Jego konstrukcja przebiega w dwóch etapach

Etap 1. W każdym wierszu (lub kolumnie) macierzy odległości szuka się elementu najmniejszego, wskazującego parę jednostek najbardziej podobnych. Otrzymane połączenie przedstawia się w postaci grafu niezorientowanego , w którym długości krawędzi są proporcjonalne do odległości pomiędzy jednostkami przyporządkowanymi poszczególnym wierzchołkom.

Etap 2. Sprawdza się spójność grafu. Jeżeli nie jest on spójny, to poszczególne jego składowe spójności (podgrafy spójne) łączą się ze sobą w miejscu wyznaczonym przez minimalną odległość pomiędzy jednostkami . wierzchołkami . należącymi do łączonych składowych. Postępowanie takie przeprowadza się dotąd, dopóki nie otrzyma się grafu spójnego, nazywanego dendrytem wrocławskim i wyznaczającego szukane uporządkowanie klasyfikowanych jednostek.

Dendryt stanowi podstawę klasyfikacji zbioru na k podzbiorów, które skupiają obiekty podobne pod względem badanych cech. Następuje to w drodze podziału dendrytu, przez odrzucenie $k-1$ najdłuższych wiązadeł. Wybór liczby k stanowi najtrudniejszy i najbardziej dyskusyjny etap analizy.
W literaturze opisywane są liczne metody prowadzące do ustalenia liczby k. W pracy wykorzystano metodę podziału .naturalnego. Aby dokonać takiego podziału należy wstępnie upo-

rządkować malejąco ciąg długości wiązadeł dendrytu kompletnego {*di*}i=1,2,...,m. Następnie zaś obliczyć indeksy:

$$w_i = \frac{\tilde{d}_{i-1}}{\tilde{d}_i}, \qquad i=2,\ldots\ldots,m$$

gdzie:

$\tilde{d}_i$ - długość wiązadeł

Wówczas za *k* przyjmuje się liczbę naturalną, dla której $W_k < W_{k+1}$. Ten podział zapewnia więc największy spadek długości wiązadeł dendrytu.
Po dokonaniu podziału zbioru wierzchołków (obiektów) na *k* podzbiorów, są podstawy do twierdzenia, że każda z tych części jest bardziej jednorodna niż cały zbiór W.
Ze względu na założenia opisanej metody, zadbać należy o to, aby zestaw zmiennych charakteryzujących obiekty poddane grupowaniu nie zawierał zmiennych skorelowanych, które są nośnikami podobnych informacji i przez to zniekształcają analizę, wywierając większy niż pozostałe cechy wpływ na dokonywane podziały. Po analizie współczynników korelacji liniowej Pearsona stwierdzono, iż para *X*1 . *X*2 jest ze sobą istotnie skorelowana (*r = .0,61*). Ze względu na to ze zbioru zmiennych diagnostycznych usunięto zmienną *X*2. Wszystkie cechy charakteryzuje dość wysoka zmienność

Po dokonaniu niezbędnej eliminacji każdy z 22 obiektów został opisany przez 5 zmiennych.
Obliczono odległości euklidesowe między każdą parą analizowanych obiektów. Macierz odległości stanowi punkt wyjścia do budowy dendrytu.
Analiza długości wiązań otrzymanego dendrytu wskazuje, iż należy odrzucić cztery najdłuższe wiązania w dendrycie spójnym i uzyskać w ten sposób pięć grup obiektów podobnych. Wiązania te zaznaczono na diagramie linią podwójną.

Elementy w ten sposób uzyskanych podzbiorów są następujące:

grupa 1: suski, wielicki, nowosądecki, Nowy Sącz, wadowicki, olkuski, limanowski, chrzanowski, myślenicki, brzeski, dąbrowski, oświęcimski,
grupa 2: gorlicki, bocheński, proszowicki, tarnowski,
grupa 3: Kraków, Tarnów, krakowski,
grupa 4: tatrzański, nowotarski,
grupa 5: miechowski.

Powiaty tworzące pierwszą, najliczniejszą grupę, obejmującą ponad połowę badanych obiektów, charakteryzują się stosunkowo niskim (na tle innych) odsetkiem bezrobotnych bez kwalifikacji zawodowych, zaś poważnym problemem jest tam brak pracy dla ludzi młodych. średnio, co czwarta osoba pozostająca bez pracy na tym obszarze, to osoba, która nie ukończyła 25 roku życia, zatem aktywne formy zwalczania bezrobocia wdrażane przez instytucje rynku pracy powinny być adresowane szczególnie do tej grupy

Skupienie drugie tworzą powiaty, w których zbiorowość bezrobotnych cechuje się bardzo wysokim udziałem osób w wieku poniżej 25 lat. Grupa ta odznacza się także stosunkowo wysokim . na tle innych . wskaźnikiem udziału osób niepełnosprawnych.

Grupę trzecią tworzą trzy powiaty, stanowiące duże skupiska miejskie i tereny wokół nich. Charakterystyczny w tej grupie jest niski odsetek ludzi młodych wśród zarejestrowanych bezrobotnych, a jednocześnie najwyższy wskaźnik osób w wieku poniżej 27 roku życia z wyższym wykształceniem. Można to tłumaczyć obecnością dużych ośrodków akademickich na tym obszarze i tym samym większym .nasyceniem. tego rynku ich absolwentami.

Grupa czwarta skupia tylko dwa powiaty. Największy problem stanowi bardzo duży wskaźnik bezrobotnych nie posiadających kwalifikacji . nie posiada ich co trzecia zarejestrowana osoba.

Inne interesujące zastosowanie **metody Taksonomii Wrocławskiej** podał Sołtysiak () do oceny i porównania stanu gospodarki dwudziestu pięciu krajów Unii Europejskiej.
W tym celu zgromadzono dla tych krajów 11 cech, które uznano za wpływające pozytywnie na gospodarkę i życie obywateli tych krajów. Do opracowania wzięto dane z roku 2001, gdyż takie były aktualnie dostępne w .Eurostat Year Book. 2004.

Wzięto pod uwagę następujące cechy:
1. liczba małżeństw zawieranych na 1000 osób,
2. udział w populacji ludności w wieku 65 lat i więcej . (w %),
3. liczba uczniów i studentów w wieku do 25 lat (bez przedszkoli) w milionach osób,
4. liczba studentów (tetrary education) . w milionach,
5. dochód narodowy na mieszkańca w PPS w cenach bieżących . w tysiącach,
6. wartość eksportu . w bilionach euro,
7. wartość importu . w bilionach euro,
8. produkcja całkowitej energii pierwotnej . w milionach,
9. zużycie energii . w milionach),
10. wielkość obszarów upraw . w milionach hektarów,
11. całkowita produkcja drewna. w milionach metrów sześciennych.

Dane te zestawiono w macierzy o wymiarach 25 x 11, o elementach x_{ij} , gdzie wskaźnik
i = 1,2,.....,25 jest wskaźnikiem kraju,
j = 1,2,.....,11 jest wskaźnikiem cechy.

Dla każdej z jedenastu cech obliczono jej wartość średnią:

$$\overline{x}_j = \frac{1}{25}\sum_{i=1}^{25} x_{ij}, \qquad j = 1,2...11$$

$$S_j = \sqrt{\frac{1}{24}\sum_{i=1}^{25}\left(x_{ij} - \overline{x}_j\right)^2}$$

Aby otrzymać jednakowe (bezwymiarowe) wartości cech, obliczono ich standaryzowane wartości:

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{S_j} \qquad i=1,2,.....,25, \qquad j=1,2,.....,11$$

obliczono .wartości. każdego z krajów pod względem rozpatrywanych jedenastu cech.

$$\gamma_i = \sum_{j=1}^{11} z_{ij}$$

.Wartości. te pokazują, że pięć pierwszych, **dominujących krajów to:**

4. Niemcy
8. Francja
25. Zjednoczone Królestwo (U.K.)
10. Włochy
19. Polska

Kraje zajmujące ostatnie pięć miejsc to:
22. Słowacja
13. Litwa
5. Estonia
12. Łotwa
21. Słowenia

Aby zbudować dendryt dwudziestu pięciu krajów na tle jedenastu cech, należy znaleźć odległości. pomiędzy tymi krajami
Za .odległość. między krajami przyjęto sumę bezwzględnych wartości różnic znormalizowanych wielkości cech

$$d_{lk} = \sum_{j=1}^{11} \left| z_{lj} - z_{kj} \right| \qquad \begin{matrix} l = 1,2...25 \\ k = 1,2,...25 \end{matrix}$$

Odległości. między krajami zestawiono w tablicy 1.

Na podstawie tablicy 1 utworzono najpierw połączenia pierwszego rzędu, to znaczy każdy kraj połączono z najbliższym. Otrzymano siedem grup krajów najbardziej do siebie zbliżonych:

I 1. Belgia ←308→ 17. Holandia

II 2. Czechy ←220→ 6. Grecja ←169→ 20. Portugalia

III 3. Dania ←294→ 9. Irlandia ←226→ 14. Luksemburg

IV 4. Niemcy ←628→ 8. Francja ←709→ 25. Zjednoczone Królestwo (U.K.)

V 10. Włochy ←619→ 7. Hiszpania ←724→ 19. Polska

22. Słowacja

↓182

VI 11. Cypr ←397→ 16. Malta ←235→ 21. Słowenia ←214→ 5. Estonia –

←108→ 13. Litwa ←140→ 12. Łotwa

VII 15. Węgry ←243→ 18. Austria ←376→ 23. Finlandia ←282→ 24. Szwecja

Rys. 1. Połączenia pierwszego rzędu dwudziestu pięciu krajów Unii

Z tych siedmiu grup utworzono połączenia drugiego rzędu, szukając najbliższych .odległości. między poszczególnymi grupami. Otrzymano dwie grupy połączeń drugiego rzędu: A i B, przedstawione na rysunku 2.

Grupę A utworzyły połączenia:

Grupa I połączyła się z Grupą VII
1. Belgia ---385--- 18. Austria

Grupa II połączyła się z Grupą VII
2. Czechy ---259 --- 15. Węgry

Grupa III połączyła się z Grupą VII
9. Irlandia ---268--- 18. Austria

Grupa VI połączyła się z Grupą VII
22. Słowacja ---227--- 15. Węgry

Grupa B powstała z połączenia Grupy IV z Grupą V:
10. Włochy ---766--- 25. Zjednoczone Królestwo (U.K.)

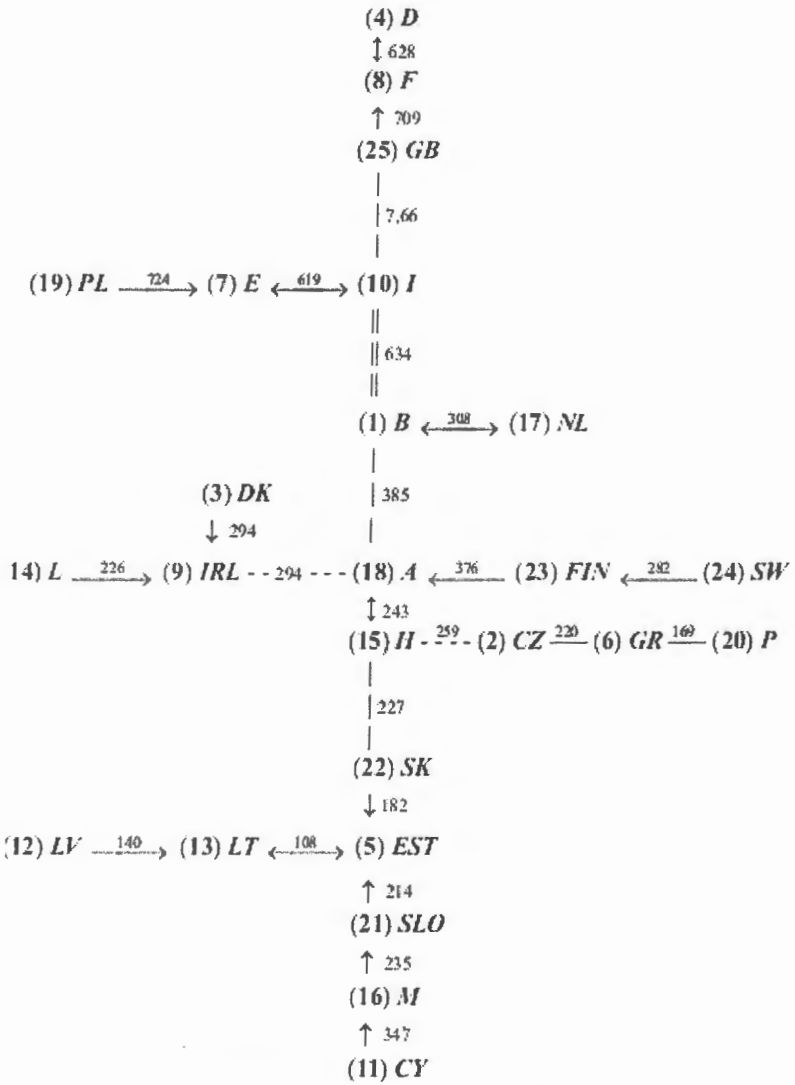
Rys. 2. Połączenia drugiego rzędu dwudziestu pięciu krajów Unii

Tablica 5. „Odległości" (x 100) między dwudziestoma pięcioma krajami na podstawie jedenastu cech

Kraj	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0																								
2	586	0																							
3	683	449	0																						
4	1937	2308	2480	0																					
5	725	335	682	2544	0																				
6	587	220	428	2410	394	0																			
7	697	739	822	1573	1132	836	0																		
8	1631	1858	2006	628	2258	1962	1134	0																	
9	1493	453	294	2426	466	394	962	1974	0																
10	634	1178	1261	1271	1415	1175	619	889	1301	0															
11	1135	825	733	3032	834	692	1462	2583	702	1805	0														
12	765	423	722	2467	205	454	1142	2268	534	1455	752	0													
13	767	375	600	2618	108	337	1106	2232	452	1363	670	140	0												
14	824	770	636	2693	577	626	1371	2173	226	1540	797	515	541	0											
15	614	259	554	2371	253	263	879	1971	487	1126	871	363	298	718	0										
16	806	440	499	2703	319	307	1123	2251	383	1476	397	358	287	462	488	0									
17	308	736	687	1907	981	703	837	1465	543	804	1197	1029	939	930	776	867	0								
18	385	359	414	2244	462	430	822	1834	268	1125	922	428	488	585	243	593	597	0							
19	1423	905	1342	1654	1006	1072	724	1082	1038	1029	1712	1216	1168	1624	995	1329	1293	1224	0						
20	498	232	425	2423	389	169	863	1983	377	1218	643	417	369	678	322	284	742	282	1111	0					
21	661	467	616	2768	214	378	1136	2282	416	1423	620	216	240	477	379	235	901	456	1362	387	0				
22	701	329	564	2574	182	300	1032	2188	398	1189	692	198	216	562	227	447	893	428	1020	333	250	0			
23	735	455	570	2174	652	519	1024	1752	492	1317	1026	662	624	779	619	657	885	376	1222	521	674	585	0		
24	563	619	646	2138	828	624	1008	1738	680	1355	1308	912	864	1001	729	979	775	434	1295	713	834	612	282	0	
25	1423	1836	1803	1103	2085	1767	1172	709	1771	766	2379	2125	1929	2180	1789	2050	1252	1699	1498	1804	2079	1995	1924	2017	0

Wreszcie poszukano najbliższego połączenia trzeciego rzędu, między państwami grupy A i B. Najkrótszym połączeniem trzeciego rzędu jest połączenie między Belgią i Włochami:
1. Belgia = = = = 10. Włochy

W ten sposób powstał dendryt dwudziestu pięciu krajów Unii. Końcową postać dendrytu można kształtować dość dowolnie, pamiętając, że rzutujemy dwadzieścia pięć punktów z przestrzeni jedenastowymiarowej na płaszczyznę.
Krajami najbardziej podobnymi do siebie (najbliższymi) są: 5. Estonia i 13. Litwa (108). Najbardziej różniącymi się krajami są: 4. Niemcy i 11. Cypr (największa odległo.ć w dendrycie . 3032). Końcową postać dendrytu przedstawiono na rysunku 3.
Rysunek 3 najlepiej unaocznia, że liderami Unii są: 4. Niemcy, 8. Francja, 25. Zjednoczone Królestwo (U.K.), outsiderami za. są: 21. Słowenia, 16. Malta i 11. Cypr.
Przeprowadzone badanie ma przede wszystkim pokazać, jak prostymi środkami można dokonywać obiektywnej oceny stanu gospodarki i znajdować uporządkowany zbiór różnic i podobieństw obiektów charakteryzujących się wieloma różnymi cechami.



Rys. 3. Dendryt dwudziestu pięciu krajów Unii na podstawie jedenastu cech (cyfry w nawiasach odpowiadają krajom podanym w tablicy 1)

3.3. METODA WARDA I METODA K-ŚREDNICH

Klasyfikację województw według ich konkurencyjności przy wykorzystaniu wybranych narzędzi analizy skupień takich jak metoda Warda oraz metoda K-średnich przedstawiły J. Wojnar oraz I. Cichocka w ().
Region jest konkurencyjny, jeżeli spełnia dwa dualne względem siebie zadania: jest w stanie przyciągnąć kapitał, szczególnie kapitał innowacyjny, oraz stwarza takie warunki zlokalizowanym na jego terenie przedsiębiorstwom, że są one w stanie wygrać konkurencję.
Polskie regiony mają silnie zróżnicowane możliwości uzyskania wysokiego poziomu konkurencyjności w globalnej gospodarce, w której trwałą przewagę konkurencyjną uzyskuje się dzięki zdolności do tworzenia innowacji. Najwyższy poziom konkurencyjności mają niewątpliwe wielkie miasta oraz otaczające je obszary . metropolia warszawska i uzyskujące cechy metropolii takie miasta, jak: Poznań, Wrocław, Kraków i Trójmiasto. W zdecydowanie mniej korzystnej sytuacji są obszary pozametropolitalne oraz regiony Polski wschodniej.
Rozwój polskich regionów dokonuje się w określonej sytuacji społecznej i gospodarczej i będąc ukierunkowanym na potrzeby i możliwości poszczególnych regionów powinien także odpowiada ć na uwarunkowania ogólnokrajowe, z których najważniejsze to:

- utrzymanie wysokiej dynamiki wzrostu gospodarczego przy dokonywaniu korzystnych zmian strukturalnych w gospodarce, polegających na zwiększaniu udziału gałęzi i branż o wysokim poziomie zaawansowania technologicznego i konkurencyjności eksportowej,
- zwiększanie zatrudnienia przez tworzenie nowych miejsc pracy i aktywne formy walki z bezrobociem,
- zwiększanie poziomu innowacyjności polskiej gospodarki w wyniku intensyfikacji i racjonalizacji badań naukowych oraz szerszego ich wykorzystywania przez przedsiębiorstwa,
- przygotowywanie kadry dla gospodarki opartej na wiedzy i rozwój społeczeństwa informacyjnego,
- rozwijanie infrastruktury transportowej i telekomunikacyjnej, zapewniającej większą spójność kraju na kilku poziomach: międzynarodowym, ogólnokrajowym, wewnątrzregionalnym i lokalnym

Część metodyczna
Przeprowadzona analiza ma dać odpowiedź na pytanie, które regiony Polski są najbardziej konkurencyjne, a więc najlepiej rozwinięte gospodarczo, o najlepszej infrastrukturze i dysponujące największym .potencjałem innowacyjności..
Skuteczną procedurą badawczą umożliwiającą uporządkowanie materiału empirycznego jest klasyfikacja sprowadzająca się do podziału zbioru obiektów na podzbiory (grupy jednostek) podobnych do siebie z punktu widzenia cech przyjętych do opisu badanego zjawiska. Podział prowadzano na podstawie podobieństwa obiektów [Grabiński i in. 1989

Zmienne losowe z tego zbioru oznaczane przez xj są zmiennymi diagnostycznymi. Do zbioru zmiennych charakteryzujących poszczególne województwa wstępnie wybrano 55 cech opublikowanych przez GUS, które pogrupowano według zagadnień: **rozwój gospodarczy, infrastruktura, innowacyjność i stan środowiska naturalnego.**
Nie wszystkie cechy okazały się jednakowo istotne z punktu widzenia przedmiotu badań i nie wszystkie powinny być w dalszej analizie uwzględniane.

Wyboru zmiennych do zbioru cech diagnostycznych dokonano posługując się kryterium

- przydatności merytorycznej w omawianej problematyce,
- zmienności . cechy diagnostyczne powinny wykazywać dostateczną zmienność przestrzenną, czyli być nośnikiem informacji różnicującej badane obiekty (województwa), w tym celu oblicza się dla analizowanych cech współczynnik zmienności, za. eliminacji podlegają te cechy, dla których współczynnik zmienności osiąga wartość mniejszą niż 0,1,
- stopnia skorelowania . zbyt silne powiązanie dwóch analizowanych cech powoduje, że są one nośnikiem podobnych informacji, dlatego przyjmuje się, że w przypadku identyfikacji zbyt wysokiej wartości współczynnika korelacji między analizowanymi cechami należy dokonać wyboru reprezentanta, zwykle kierując się przesłankami merytorycznymi; za progowy poziom współczynnika korelacji przyjmuje się zazwyczaj r = 0,7 [Nowak 1990].

Przeprowadzenie normalizacji zmiennych zapewnia eliminację wymienionych ograniczeń i trudności interpretacyjnych

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_{x_j}}$$

W przypadku standaryzacji następuje wyrównanie dyspersji oraz poziomu wartości cechy, ponieważ wariancje zmiennych diagnostycznych są równe 1, a średnie arytmetyczne 0, w związku z tym każda ze zmiennych w jednakowym stopniu wpływa na końcowe wyniki prowadzonej analizy.
W wyniku zastosowania wyżej opisanej procedury, ostatecznie otrzymano zbiór 35 cech diagnostycznych, które w sposób możliwie pełny charakteryzują badane województwa.

Ze względu na charakteryzujące je cechy wybrano ostatecznie do **analizy metodę Warda**, której wyniki porównano z **metodą k-.średnich**. Efektem tej analizy jest uzyskanie segmentów jednolitych wewnętrznie i różniących się między sobą.

Metoda aglomeracji Warda należy do grupy metod hierarchicznych. Przyjmuje ona za punkt wyjścia maksymalną liczbę skupień, równą liczbie badanych obiektów (skupienia jednoelementowe). Algorytmy grupowania mają następnie na celu łączenie kolejno obiektów o rosnącej odmienności. Metoda ta zmierza do minimalizacji sumy kwadratów odchyleń wewnątrz skupień. Na każdym etapie spośród wszystkich możliwych do łączenia par skupień wybiera się tę, która w rezultacie łączenia daje skupienie o minimalnym zróżnicowaniu. Miarą tego zróżnicowania względem wartości średnich jest wyrażenie ESS, zwane też błędem sumy kwadratów, które określane jest wzorem:

$$ESS = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

gdzie:

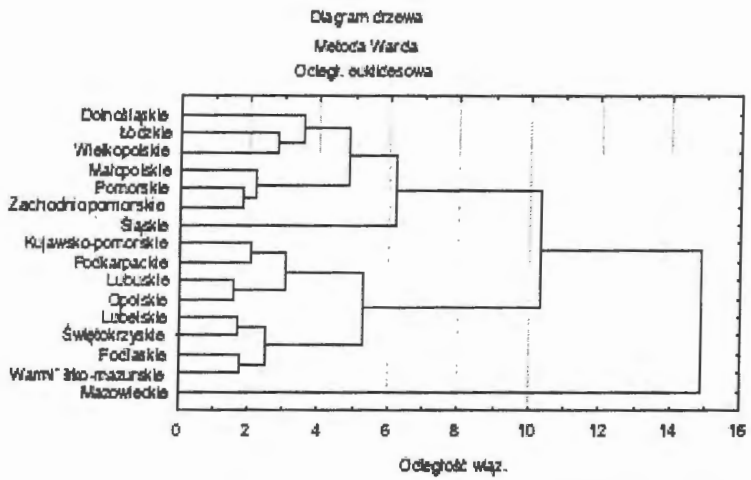
x_i - wartość zmiennej będącej kryterium segmentacji dla i-tego obiektu,
n - liczba obiektów w skupieniu.

Empirycznie metoda daje bardzo dobre wyniki (grupy są bardzo homogeniczne), jednak ma skłonność do tworzenia skupień o małej wielkości i o podobnych rozmiarach. Obok wysokiej efektywności do zalet omawianej metody należy także zaliczyć sposób prezentacji wyników analizy za pomocą dendrogramu. Taka wizualizacja pozwala szybko zorientować się w strukturze zbiorowości. Istnieje jednak pewna dowolność w interpretacji wielkości i liczby skupień.

Metoda k-średnich różni się od metody aglomeracji Warda. Wymaga sformułowania a priori założenia liczby skupień uzyskanych w wyniku grupowania. Grupowanie metodą k-średnich jest procedurą iteracyjną, w każdej iteracji pewne obiekty są przenoszone do innych skupień, celem uzyskania minimalnego zróżnicowania wewnątrzgrupowego i maksymalnego międzygrupowego. Konieczność określenia liczby skupień uznawana jest za wadę tej metody, o ile struktura zbioru obiektów nie jest znana i nie ma możliwości sformułowania hipotez odnośnie oczekiwanej liczby skupień. Zaletą metody jest to, że gwarantuje otrzymanie k-skupień w możliwie największym stopniu różniących się od siebie.

Celem pogrupowania województw ze względu na **potencjał gospodarczy** analizę przeprowadzono na podstawie następujących zmiennych X1-X13
X1 Przeciętne miesięczne wynagrodzenie brutto [zł]
X2 Przeciętny miesięczny dochód na 1 osobę w gospodarstwach domowych [zł]
X3 Wydatki budżetów województw [zł/1 mieszkańca]
X4 Nakłady inwestycyjne (ceny bieżące) [zł/1 mieszkańca]
X5 Wartość dodana brutto na 1 pracującego [zł]
X6 PKB [zł/1 mieszkańca]
X6 Produkcja sprzedana przemysłu [mln zł]
X7 Wartość dodana brutto według rodzajów działalności: rolnictwo, łowiectwo, leśnictwo
X8 Wartość dodana brutto według rodzajów działalności: usługi rynkowe
X9 Stopa bezrobocia [%]
X10 Bezrobotni zarejestrowani [tys.]
X11 Pracujący [tys.]
X12 Wskaźnik cen towarów i usług konsumpcyjnych
X13 Liczba ofert pracy [tys.]

W wyniku przeprowadzenia grupowania metodą Warda otrzymano dendrogram



W zależności od rozpatrywanej odległości między obiektami wyróżniamy 3 lub 4 bardziej jednorodne skupienia województw:
– skupienie I: mazowieckie,
– skupienie II: śląskie,
– skupienie III: dolnośląskie, łódzkie, wielkopolskie, małopolskie, pomorskie, zachodniopomorskie,
– skupienie IV: kujawsko-pomorskie, podkarpackie, lubuskie, opolskie, lubelskie, świętokrzyskie, podlaskie, warmińsko-mazurskie.

Wyniki uzyskane metodą Warda są zgodne z klasyfikacją województw uzyskaną metodą k-średnich, przy założeniu czterech skupień. Zakładając podział na trzy skupienia różnica polega na tym, że województwo mazowieckie znajduje się w jednej grupie z województwem śląskim. Analizując wartości średnie w poszczególnych grupach można zaobserwować, że skupienie pierwsze, do którego należy tylko województwo mazowieckie, wyraźnie odbiega od pozostałych, szczególnie pod względem takich cech, jak: wartość PKB . 40 817zł/1 mieszkańca (średnia dla kraju . 25 767 zł/ 1 mieszkańca), przeciętne miesięczne wynagrodzenie brutto . 3166,02 zł (średnia dla kraju .2475,88 zł). W województwie tym można zaobserwować bardzo dobrą sytuację na rynku pracy, stopa bezrobocia na koniec 2006 roku wynosiła 11,9% (w kraju 14,9%). W województwie mazowieckim utrzymuje się bardzo wysokie tempo wzrostu gospodarczego, stabilna polityka gospodarcza, której celem jest tworzenie warunków do rozwoju przedsiębiorczości, wyraźnie widoczny jest rozwój nowoczesnych sektorów gospodarki i alokacja inwestycji zagranicznych.
Kolejnym skupieniem jednoelementowym jest województwo śląskie, które po województwie mazowieckim stanowi najbogatszy regionem w Polsce. Wytwarza 14% PKB Polski, a średnia pensja w porównywanym okresie wynosiła 2560,33 zł. Stopa bezrobocia kształtowała się na

poziomie 9,3%. Główną gałęzią gospodarki województwa śląskiego są usługi. Niemniej jednak bardzo duży udział ma także przemysł, który wytwarza około 34% PKB. Najmniej ludności pracuje w rolnictwie i leśnictwie. Najważniejsze gałęzie to górnictwo, hutnictwo oraz produkcja energii elektrycznej. Województwo wytwarza 92% węgla kamiennego w Polsce, 83% samochodów oraz 70% stali surowej.
Skupienie trzecie stanowi pięć województw, dla których wartości badanych cech oscylują w okolicach średniej krajowej. Najniższy poziom rozwoju gospodarczego cechuje osiem województw należących do ostatniego skupienia. Są to głównie regiony położone we wschodniej części kraju, województwa lubelskie, podkarpackie, warmińsko-mazurskie, podlaskie i świętokrzyskie, oraz województwa kujawsko-pomorskie, opolskie i lubuskie. Średnia wartość PKB na 1 mieszkańca w tej grupie wynosiła 20 175,14 z^3, a przeciętne miesięczne wynagrodzenie kształtowało się na poziomie 2150,10 z^3.

Istotnym czynnikiem kształtującym atrakcyjność regionu jest **poziom rozwoju infrastruktury**.
Analizując infrastrukturę województw w oparciu o cechy X14-X24 otrzymano bardzo zbliżony podział skupień. Do skupienia I charakteryzującego się najwyższymi wartościami badanych cech należy województwo śląskie i mazowieckie.

X14 Zasoby mieszkaniowe na 1000 ludności
X15 Mieszkańcy korzystający z wodociągu (% ludności ogółem)
X16 Mieszkańcy korzystający z sieci kanalizacyjnej (% ludności ogółem)
X17 Gospodarstwa domowe podłączone do sieci gazowej (% gospodarstw ogółem)
X18 Mieszkańcy obsługiwani przez oczyszczalnie .cieków (% ludności ogółem)
X19 Eksploatowane linie kolejowe na 100 km2 [km]
X20 Drogi publiczne o nawierzchni twardej na 100 km2 [km]
X21 Łącza główne wszystkich operatorów telefonii przewodowej na 1000 mieszkańców [szt]
X22 Dostępność łóżek w szpitalach w przeliczeniu na 10 tys.
X23 Liczba lekarzy przypadająca na 1000 mieszkańców (lekarzy pracujących w zawodzie)
X24 Studenci szkół wyższych na 10 tys. Ludności

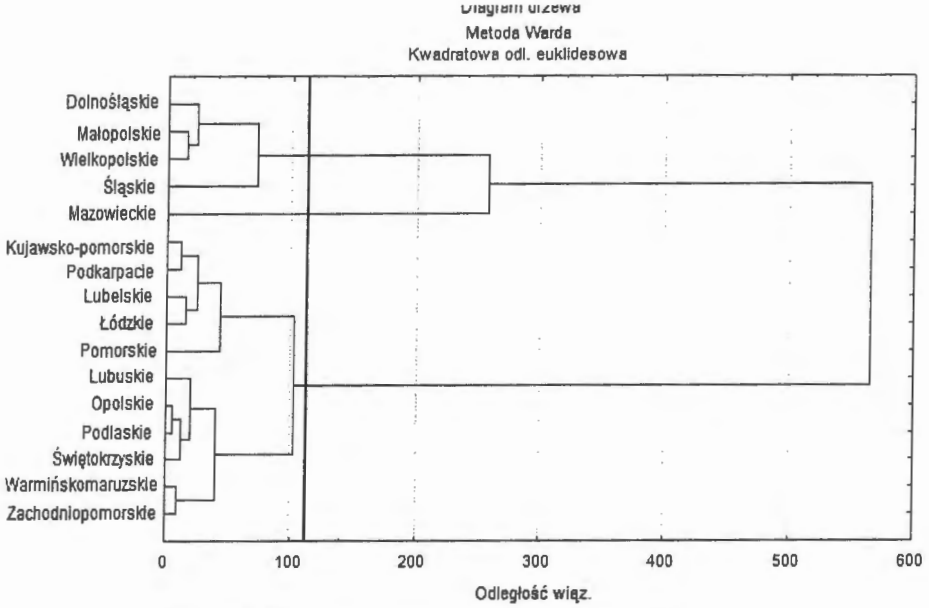
Kluczem do uzyskania przewagi konkurencyjnej regionu są innowacje i nowe technologie. Zwiększenie innowacyjności i konkurencyjności w krajach UE stało się czołowym priorytetem wspólnej polityki spójności. W przygotowanym przez Komisję Europejską raporcie dotyczącym innowacyjności [European ... 2006] w Unii Europejskiej Polska sklasyfikowana jest na 21 pozycji. Warto podkreślić, że niskie nakłady na badanie i rozwój, niskie zatrudnianie w branżach wysokich technologii, brak powiązań sfery naukowo-badawczej z biznesem to słabe punkty polskiej gospodarki. Poziom innowacyjności danego kraju stanowi wypadkową poziomu innowacyjności poszczególnych jego regionów.
Wykorzystując zmienne X25- X45, przeprowadzono klasyfikację województw ze względu na **potencjał innowacyjności**..

X25 Liczba jednostek zajmujących się działalnością B+R
X26 Udział osób zatrudnionych w działalności B+R [%]
X27 Nakłady na działalność B+R [mln zł]
X28 Nakłady bieżące na działalność B+R [mln zł]
X29 Nakłady z budżetu państwa na działalność B+R [%]
X30 Nakłady podmiotów gospodarczych na B+R [%]

X31 Nakłady na działalność innowacyjną w przemyśle [%]
X32 Nakłady na działalność innowacyjną w przemyśle [z]ł
X33 Nakłady na zakup gotowej technologii w postaci dokumentacji i praw w przemyśle [mln zł]
X34 Nakłady na budynki i budowle oraz grunty w przemyśle [mln zł]
X35 Nakłady na maszyny, urządzenia techniczne i narzędzia oraz środki transportu [zł]
X36 Nakłady na szkolenie personelu związane z działalnością innowacyjną w [zł]
X37 Nakłady na marketing dotyczący nowych lub istotnie ulepszonych produktów [zł]
X38 Liczba linii produkcyjnych zainstalowanych w przemyśle [szt]
X39 Udział przedsiębiorstw przemysłowych, które prowadziły działalność innowacyjną [%]
X40 Liczba zgłoszonych wynalazków krajowych na 100 tys. mieszkańców.
X41 Liczba osób zatrudnionych w działalności badawczo rozwojowej na 1 tys. mieszkańców.
X42 Udział osób uczących się i dokształcających wśród ogółu ludności w wieku 25-64 [%]
X43 Wydatki przedsiębiorstw na badania i rozwój w % PKB
X44 Ludność z wykształceniem wyższym (% populacji w wieku 25-64)
X45 Pracujący w wysoko zaawansowanych technologiach (% ogółu pracowników)

W efekcie prowadzonej analizy przy wykorzystaniu metod klasyfikacji pogrupowano województwa na trzy skupienia przedstawione na poniższych rysunkach.

Rysunek 5. Wyniki grupowania województw metodą k-średnich



Rysunek 6. Wyniki grupowania województw metodą Warda

Należy podkreślić, że taką hierarchizację województw wykazywały niemal wszystkie metody klasyfikacji obiektów. Do grupy obiektów najwyżej sklasyfikowanych należy województwo mazowieckie, które spośród 20 cech reprezentujących rozwój innowacyjności regionu w przypadku 18 posiadało najkorzystniejsze wartości. Cechami, która zdecydowanie wyróżniają to województwo od pozostałych jest odsetek osób zatrudnionych w działalności badawczo-rozwojowej, wartość ta kształtuje się na poziomie 33,1%, oraz nakłady na działalność badawczo rozwojową.

Bardzo istotnym elementem konkurencyjności regionu jest zanieczyszczenie **środowiska naturalnego**. Obiekty zostały pogrupowane względem cech X46-X55.

X46 Emisja zanieczyszczeń powietrza gazowych (bez CO2) [tys. ton]
X47 Emisja zanieczyszczeń pyłowych z zakładów szczególnie uciążliwych [tys. ton/rok]
X48 Emisja zanieczyszczeń gazowych z zakładów szczególnie uciążliwych[tys. ton/rok]
X49 ścieki przemysłowe nie oczyszczane odprowadzane bezpośrednio z zakładów przemysłowych [w m3 /100km2]
X50 ścieki przemysłowe i komunalne wymagające oczyszczenia [m3/100 km2]
X51 Odpady przemysłowe wytworzone w ciągu roku [tys. ton]
X52 Odpady (z wyłączeniem komunalnych) wytworzone na 1 km2 w tonach
X53 Zakłady szczególnie uciążliwe emitujące zanieczyszczenia powietrza według wielkości emisji zanieczyszczeń pyłowych
(liczba zakładów/100 km2)

X54 Zakłady odprowadzające .cieki do sieci kanalizacji miejskiej bez oczyszczania (jaki % ogółu zakładów)
X55 ścieki odprowadzane siecią kanalizacji - nieoczyszczone [%]

Wszystkie metody grupowania wyłoniły jako pierwsze, skupienie jednoelementowe, które tworzy województwo śląskie. Województwo to zajmuje pozycję lidera w rankingu ze względu na bardzo wysokie wartości analizowanych cech.
Biorąc pod uwagę fakt, że obszar województwa śląskiego zajmuje jedynie 3,9% powierzchni kraju, ilość substancji pogarszających stan środowiska oraz wytwarzanych w regionie odpadów jest nieporównywalna z jakimkolwiek innym regionem w Polsce.
Podział na drugie i trzecie skupienie jest nieco inny w przypadku wybranych metod grupowania.
Województwa dolnośląskie, mazowieckie i wielkopolskie należą do tego samego skupienia i wartości średnie analizowanych cech dla tych województw są na poziomie wyższym niż średnia dla kraju.
Trzy województwa zachodniopomorskie, łódzkie i małopolskie w zależności od metody grupowania znalazły się w różnych skupieniach. Skupienie trzecie koncentruje województwa, dla których poziom cech jest zbliżony do średniej krajowej. Są to województwa kujawsko-pomorskie, lubelskie, podkarpackie i pomorskie.
Ostatnie skupienie, do którego niezależnie od metody grupowania zaliczonych zostało 5 województw o najmniejszym stopniu zanieczyszczenia środowiska, tworzą województwa: lubuskie, opolskie, podlaskie, świętokrzyskie, warmińsko-mazurskie.

Tabela 1. Grupowanie województw według stopnia zanieczyszczenia środowiska naturalnego

Metody grupowania	Skupienie I	Skupienie II	Skupienie III	Skupienie IV
Warda, średnich połączeń, pełnego wiązania	śląskie	dolnośląskie, mazowieckie, wielkopolskie, łódzkie, małopolskie,	kujawsko-pomorskie, lubelskie, podkarpackie, pomorskie, zachodnio-pomorskie	lubuskie, opolskie, podlaskie, świętokrzyskie, warmińsko-mazurskie
Metoda *k*-średnich	śląskie	dolnośląskie, mazowieckie, wielkopolskie, zachodniopomorskie	kujawsko-pomorskie, lubelskie, łódzkie, małopolskie, podkarpackie, pomorskie	lubuskie, opolskie, podlaskie, świętokrzyskie, warmińsko-mazurskie

W czołówce województw o najwyższej atrakcyjności inwestycyjnej znajdują się województwa śląskie, mazowieckie i dolnośląskie. Najbardziej dynamiczne rozwijające się województwa: mazowieckie, śląskie, wielkopolskie i dolnośląskie cechują się jednocześnie największym potencjałem gospodarczym.
Ponad połowa produktu krajowego brutto powstającego w Polsce pochodzi z tych 4 województw, w tym 1/5 z mazowieckiego. Ranking konkurencyjności zamykają województwa: podkarpackie, świętokrzyskie, lubelskie, podlaskie i opolskie. Regiony te we wszystkich analizowanych kategoriach należą do ostatniego skupienia.
Należy podkreślić, że na zróżnicowanie szans rozwoju poszczególnych obszarów Polski istotny wpływ mają czynniki historyczne, a także międzynarodowe relacje regionów przygranicz-

nych. Przez stulecia wschodnia część kraju była słabiej rozwinięta i gorzej wyposażona materialnie, niż obszary położone na zachód od Wisły

4. LITERATURA

1. Badach E., 2007: Zastosowanie taksonomii wrocławskiej w badaniach populacji osób bezrobotnych w województwie małopolskim, Stowarzyszenie Ekonomistów Rolnictwa i Agrobiznesu, Roczniki Naukowe, tom IX, z. 3, s.11-14

2. Grabiński T. 1992: Metody taksonometrii. AE, Kraków.

3. Grabiński T., Wydymus S., Zeliaś A. 1989: Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych. PWN, Warszawa.

4. Gorzelak G. 2005: Weryfikacja struktury celów, priorytetów, oraz kierunków działań narodowej strategii rozwoju regionalnego na lata 2007-2013,

5. Górka J., Chmurka M., 2004: Wykorzystanie wybranej metody taksonomicznej do klasyfikacji powiatów województwa wielkopolskiego na podstawie wybranych cech turystyczno-rekreacyjnych , Roczniki Naukowe AWF w Poznaniu, Poznań, z. 53 , s.99-108

6. Hellwig Z. 1968: Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr. Przegląd Statystyczny, nr 4, s.307-324.

7. Nowak K., 1990, Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych. PWE, Warszawa.

8. Ostasiewicz W. 1998: Statystyczne metody analizy danych. AE, Wrocław.

9. Sołtysiak J., 2006: Zastosowanie Taksonomii Wrocławskiej do oceny stanu gospodarki dwudziestu pięciu krajów Unii Europejskiej, Studia Gdańskie. Wizje i rzeczywistość, t. III, s.43 – 52.

10. Wojnar J., Cichocka M., 2007: Klasyfikacja województw według ich konkurencyjności przy wykorzystaniu wybranych narzędzi analizy skupień, Stowarzyszenie Ekonomistów Rolnictwa i Agrobiznesu, Roczniki Naukowe, tom XI, z. 2, s. 278- 284

11. Taksonomia – Klasyfikacja i analiza danych – teoria i zastosowania, Prace Naukowe Akademii Ekonomicznej im. Oskara Lanego we Wrocławiu, Wydawnictwo Akademii Ekonomicznej im. Oskara Lanego we Wrocławiu, Wrocław, 2004 i 2005

12. Zając K., 1988, Zarys metod statystycznych. PWE, Warszawa.