

## STUDIA METODOLOGICZNE

**Małgorzata MACHOWSKA-SZEWCZYK, Agnieszka SOMPOLSKA-  
-RZECHUŁA**

### Dynamiczny dobór cech w taksonomicznej analizie obiektów

---

Zmienność zjawisk oraz ich wahania spowodowane działaniem czynników losowych powodują, że wraz z upływem czasu mogą zmieniać się relacje pomiędzy poszczególnymi obiektami. Uwzględnianie tych zmian i wahań ma szczególne znaczenie przy prognozowaniu zjawisk i powinno stanowić podstawę każdej decyzji dotyczącej badanych obiektów. Klasyczne metody doboru zmie-

nnych są mniej przydatne w przypadku posługiwania się danymi przekrojowo-czasowymi. Współczynniki zmienności i korelacji liniowej obliczane na podstawie danych dotyczących ostatniego okresu, a w przypadku relatywnie dużych wahań wartości zmiennych diagnostycznych — na podstawie kilku ostatnich okresów, nie mogą znaleźć zastosowania do wyboru zmiennych. W miarę upływu czasu często następują zmiany wartości cech, ich natężenia, a także kierunku powiązań między zmiennymi diagnostycznymi. Różne zmienne mogą wywierać istotny wpływ na badane zjawisko. Działanie jednych zmiennych może się nasilać, innych słabnąć. Zatem warto zastosować nieco inne postępowanie, które uwzględniłoby zmiany zachodzące w powiązaniach między przyjętymi zmiennymi i umożliwiłoby lepsze zrozumienie procesów dynamicznych<sup>1</sup>.

Zasadniczym celem artykułu jest prezentacja wyników badania taksonomicznego, w którym wykorzystano dynamiczne podejście na etapie doboru cech. Otrzymany zbiór cech diagnostycznych zastosowano do klasyfikacji średniego poziomu rozwoju społeczno-gospodarczego obserwowanego w województwach w latach 2002—2009.

## *DYNAMICZNY DOBÓR CECH DIAGNOSTYCZNYCH<sup>2</sup>*

Dany jest zbiór obiektów  $\{O_1, O_2, \dots, O_n\}$  oraz zbiór zmiennych diagnostycznych  $\{X_1, X_2, \dots, X_m\}$ . Macierz obserwacji  $\mathbf{X}^t$  ( $t=1, \dots, T$ ) ma postać:

$$\mathbf{X}^t = \begin{bmatrix} x_{11}^t & x_{12}^t & \dots & x_{1m}^t \\ x_{21}^t & x_{22}^t & \dots & x_{2m}^t \\ \dots & \dots & \dots & \dots \\ x_{n1}^t & x_{n2}^t & \dots & x_{nm}^t \end{bmatrix} \quad (1)$$

gdzie  $x_{ij}^t$  — zaobserwowana wartość zmiennej diagnostycznej  $X_j$  ( $j=1, \dots, m$ ) w obiekcie  $O_i$  ( $i=1, \dots, n$ ) w okresie  $t$  ( $t=1, \dots, T$ ).

Z potencjalnej listy zmiennych diagnostycznych eliminuje się te, które wykazują niską zmienność czasowo-przestrzenną. Podstawą eliminacji zmiennych diagnostycznych może być współczynnik zmienności  $v_j^t$  ( $j=1, \dots, m$ ;  $t=1, \dots, T$ ), który uwzględniając zróżnicowanie wartości zmiennych diagnostycznych  $X_1, X_2, \dots, X_m$  w przekroju obiektów  $O_i$  ( $i=1, \dots, n$ ) w kolejnych okresach, można zapisać następująco:

<sup>1</sup> Zeliaś (2000).

<sup>2</sup> Opracowano na podstawie Grabiński i in. (1989).

$$v_j^t = \frac{s_j^t}{|\bar{x}_j^t|} \quad \bar{x}_j^t \neq 0 \quad (j=1, \dots, m; \quad t=1, \dots, T) \quad (2)$$

gdzie:

$$\bar{x}_j^t = \frac{\sum_{i=1}^n x_{ij}^t}{n} \quad s_j^t = \sqrt{\frac{\sum_{i=1}^n (x_{ij}^t - \bar{x}_j^t)^2}{n}} \quad (3)$$

przy czym  $s_j^t$  oznacza odchylenie standardowe, zaś  $\bar{x}_j^t$  jest średnią arytmetyczną zmiennej  $X_j$  w okresie  $t$ .

Macierz współczynników zmienności jest następująca:

$$\mathbf{V} = \begin{bmatrix} v_1^1 & v_2^1 & \dots & v_m^1 \\ v_1^2 & v_2^2 & \dots & v_m^2 \\ \dots & \dots & \dots & \dots \\ v_1^T & v_2^T & \dots & v_m^T \end{bmatrix} \quad (4)$$

Otrzymuje się w ten sposób ciąg współczynników  $v_j^t (t=1, \dots, T; \quad j=1, \dots, m)$  w kolejnych latach. Dla każdej zmiennej  $X_j$  oblicza się średnią arytmetyczną współczynników zmienności:

$$\bar{v}_j = \frac{\sum_{t=1}^T v_j^t}{T} \quad (j=1, \dots, m) \quad (5)$$

Jeżeli wartości  $v_j^t$  układają się w postaci poziomej wstęgi, czyli oscylują w sposób przypadkowy wokół  $\bar{v}_j$ , to znaczy, że współczynniki zmienności  $v_j^t (t=1, \dots, T; \quad j=1, \dots, m)$  nie zmieniają się i odpowiadające im cechy podlegają eliminacji. W przeciwnym przypadku można wnioskować, że  $v_j^t$  nie są stałe, lecz mają tendencję do wzrostu w miarę upływu czasu.

Jeżeli ciągi wartości  $v_j^t (t=1, \dots, T; \quad j=1, \dots, m)$  wykazują wyraźne systematyczne zmiany, można wyznaczyć ich trendy. W taksonomicznych badaniach przestrzenno-czasowych zazwyczaj rozpatruje się niezbyt długie przedziały czasu, dlatego zmiany współczynników mogą wykazywać trendy liniowe:

$$v_j^t = \beta_{0j} + \beta_{1j}t + u_{tj} \quad (t=1, \dots, T; \quad j=1, \dots, m) \quad (6)$$

gdzie:

$\beta_{0j}, \beta_{1j}$  — parametry funkcji trendu liniowego,

$t$  — zmienna czasowa,

$u_{tj}$  — składnik losowy.

Parametry  $\beta_{0j}, \beta_{1j}$  funkcji trendu liniowego są estymowane za pomocą klasycznej metody najmniejszych kwadratów. Znajduje się tu oceny parametrów funkcji trendu określonej wzorem:

$$\hat{v}_j^t = b_{0j} + b_{1j}t \quad (t=1, \dots, T; j=1, \dots, m) \quad (7)$$

Występowanie trendu liniowego można sprawdzić stosując test na jego istnienie. Jeżeli trend jest liniowy, to można zweryfikować hipotezę  $H_0: \beta_{1j} = 0$ . Jako hipotezę alternatywną przyjmuje się  $H_1: \beta_{1j} \neq 0$ . Jeżeli parametry  $\beta_{1j}$  ( $j=1, \dots, m$ ) różnią się istotnie od zera, to obliczając prognozę zmiennej  $v_j^t$  na przyszły okres  $T'$  należy ekstrapolować relację  $v_j^t = b_{0j} + b_{1j}t + u_{tj}$  ( $t=1, \dots, T; j=1, \dots, m$ ), przyjmując, że w okresie prognozowanym  $T'$  współczynniki zmienności będą równe:

$$\hat{v}_j^{T'} = b_{0j} + b_{1j}T' \quad (T'=n+1, n+2, \dots; j=1, \dots, m) \quad (8)$$

Eliminacji podlegają zmienne diagnostyczne, dla których spełniona jest nierówność  $\hat{v}_j^{T'} \leq \varepsilon$ , gdzie  $\varepsilon$  jest arbitralnie zadaną, dostatecznie małą liczbą dodatnią, np.  $\varepsilon = 0,35$ . Można przypuszczać, że w przyszłości nie będą one odznaczać się dostateczną zmiennością czasową.

W miarę upływu czasu może zmieniać się nie tylko zróżnicowanie wartości zmiennych diagnostycznych  $X_1, X_2, \dots, X_m$  w przekroju badanych obiektów, lecz zmianie mogą ulec także powiązania między nimi. Wówczas należy wyznaczyć na podstawie macierzy obserwacji  $\mathbf{X}^t (t=1, \dots, T)$  współczynniki korelacji liniowej dla kolejnych okresów.

Współczynniki te utworzą następujące kwadratowe macierze korelacji:

$$\mathbf{R}^t = \begin{bmatrix} 1 & r_{12}^t & \cdots & r_{1m}^t \\ r_{21}^t & 1 & \cdots & r_{2m}^t \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1}^t & r_{m2}^t & \cdots & 1 \end{bmatrix} \quad (9)$$

gdzie  $r_{ij}^t$  ( $i, j=1, \dots, m; i \neq j; t=1, \dots, T$ ) jest współczynnikiem korelacji liniowej między zmiennymi  $X_i, X_j$  obliczonymi na podstawie danych z okresu  $t$ .

W rezultacie otrzymuje się obraz dynamicznych zmian zachodzących w sile powiązań między rozpatrywanymi zmiennymi. Do wykazania natężenia i kierunku tych zmian w badanym przedziale czasu  $\langle 1, T \rangle$  da się wykorzystać funkcje trendu współczynników korelacji  $r_{ij}^t$  ( $t=1, \dots, T$ ). Można przyjąć, że będą to trendy liniowe postaci:

$$r_{ij,t} = \alpha_{0,ij} + \alpha_{1ij}t + u_{ij,t} \quad (i, j=1, \dots, m; i \neq j; t=1, \dots, T) \quad (10)$$

Ważne jest, czy wartości współczynników  $r_{ij}^t$  rosną czy też mają charakter stacjonarny. Jeśli okaże się, że parametr  $\alpha_{1ij}$  różni się istotnie od zera, to zmiany w sile powiązań zmiennych  $X_i$  i  $X_j$  w przedziale czasu  $\langle 1, T \rangle$  uznaje się za znamienne. Jeżeli natomiast parametr  $\alpha_{1ij}$  okaże się statystycznie nieistotny, to zmiany w sile powiązań zmiennych  $X_i$  i  $X_j$  w badanym przedziale czasowym można uznać za losowe. Uwzględnienie tych istotnych zmian w procedurze wyboru zmiennych diagnostycznych sprawia, że otrzymany zestaw zmiennych końcowych będzie nie tylko aktualny w okresie badanym, ale także w przyszłości.

Oszacowane funkcje trendu:

$$\hat{r}_{ij,T'} = a_{0,ij} + a_{1ij}T' \quad (T' = n+1, n+2, \dots; i, j=1, \dots, m; i \neq j) \quad (11)$$

są podstawą do wyznaczenia przyszłych wartości poszczególnych współczynników korelacji liniowej. Otrzymuje się macierz prognoz współczynników korelacji liniowej, która ma postać:

$$\mathbf{R}^{T'} = \begin{bmatrix} 1 & \hat{r}_{12}^{T'} & \dots & \hat{r}_{1m}^{T'} \\ \hat{r}_{21}^{T'} & 1 & \dots & \hat{r}_{2m}^{T'} \\ \dots & \dots & \dots & \dots \\ \hat{r}_{m1}^{T'} & \hat{r}_{m2}^{T'} & \dots & 1 \end{bmatrix} \quad (T' = T+1, T+2, \dots) \quad (12)$$

Korzystając z macierzy  $\mathbf{R}^{T'}$  można dokonać redukcji zbioru zmiennych potencjalnych  $X_1, X_2, \dots, X_m$  za pomocą wybranej metody.

Jedną z propozycji może być metoda doboru zmiennych za pomocą macierzy odwrotnej (Malina, Zeliaś, 1997; Młodak, 2006), w której procedura eliminacji

jest następująca: korzystając z macierzy współczynników korelacji<sup>3</sup>  $\mathbf{R}$  wyznacza się macierz odwrotną  $\mathbf{R}^{-1}=[r_{ij}]$ . Element diagonalny  $r_{ii}$  macierzy  $\mathbf{R}^{-1}$  jest równy jedności, jeśli zmienna  $X_i$  jest ortogonalna względem pozostałych zmiennych. W przypadku nieortogonalności  $r_{ii} \in (1, +\infty)$ . Gdy zmienna jest nadmiernie skorelowana z pozostałymi, wówczas element diagonalny macierzy odwrotnej  $\mathbf{R}^{-1}$  jest znacznie większy od jedności, co jest symptomem złego uwarunkowania macierzy  $\mathbf{R}$ , czyli macierz odwrotna nie jest stabilna<sup>4</sup>.

Cechy nadmiernie skorelowane, którym odpowiadają elementy diagonalne  $r_{ii}$  o wartościach większych niż 10, są eliminowane z pierwotnego zbioru cech. Jeżeli takie elementy nie występują, to procedurę uznaje się za zakończoną. W przeciwnym przypadku ponownie wyznacza się macierz odwrotną dla zredukowanego zbioru cech i analizuje jej elementy diagonalne. Procedurę powtarza się do momentu osiągnięcia stabilności macierzy  $\mathbf{R}^{-1}$ , czyli pojawienia się elementów diagonalnych, których wartości nie przekraczają znacząco 10.

### KLASYFIKACJA Z WYKORZYSTANIEM ODLEGŁOŚCI PENROSE'A

W artykule przedstawiono technikę klasyfikacji wykorzystującą odległości między populacjami, w której dobór cech przeprowadzono z uwzględnieniem trendów liniowych współczynników zmienności oraz korelacji. Wygodna do oceny odległości między dwoma momentami obserwacji wartości ustalonych zmiennych przyjętych w badaniu obiektów jest odległość Penrose'a, którą wyznacza się następująco (Ostasiewicz, 1998; Strahl, 2006):

$$d_{tt'} = \frac{1}{m} \sum_{j=1}^m \frac{(\bar{x}_j^t - \bar{x}_j^{t'})^2}{s_j^2} \quad (13)$$

gdzie:

$\bar{x}_j^t, \bar{x}_j^{t'}$  — wartości średnie zmiennej  $X_j$  w momentach  $t, t'$ ,

$s_j^2$  — wariancja zmiennej  $X_j$  wyznaczona zgodnie ze wzorem:

$$s_j^2 = \frac{1}{n \cdot T} \sum_{t=1}^T \sum_{i=1}^n (x_{ij}^t - \bar{x}_j^t)^2 \quad (14)$$

<sup>3</sup> W analizowanym przypadku jest to macierz prognoz współczynników korelacji  $\mathbf{R}^T$ .

<sup>4</sup> Stabilność macierzy odwrotnej oznacza, że małym zmianom w elementach danej macierzy odpowiadają małe zmiany w elementach macierzy odwrotnej. Jakość uwarunkowania macierzy określa się za pomocą tzw. wskaźników uwarunkowania macierzy, np. liczby Turinga, liczby Todda (Malina, Zeliaś, 1997).

Miara odległości Penrose’a uwzględnia jedynie wariancje poszczególnych zmiennych w całym badanym okresie, pomijając wpływ wzajemnego skorelowania poszczególnych zmiennych na odległość między populacjami w dwóch wyróżnionych momentach. Jeśli pośród badanych cech znajdują się silnie skorelowane, to ich wpływ na wartość miary Penrose’a jest większy niż pozostałych cech, dlatego ważne jest, aby wyznaczanie macierzy odległości między populacjami według wzoru (13) poprzedzić eliminacją zmiennych silnie skorelowanych między sobą. Wyznaczona macierz odległości Penrose’a może stanowić punkt wyjścia do klasyfikacji obiektów przeprowadzonej wybraną metodą hierarchiczną.

Jedną z najczęściej stosowanych metod hierarchicznych jest metoda Warda, która na każdym etapie usiłuje optymalizować otrzymany podział przez połączenie dwóch elementów, stosując kryterium minimalnego wzrostu łącznej wewnątrzgrupowej sumy kwadratów odchyleń wszystkich wartości zmiennych dla każdego obiektu od ich średnich grupowych, pod warunkiem że punktem wyjścia do procesu aglomeracji jest macierz kwadratów odległości euklidesowych. Zapewnia zatem homogeniczność wewnątrz skupień i heterogeniczność między skupieniami, przez co uznawana jest za bardzo efektywną (Ward, 1963). Metoda Warda jest prawie o 40% bardziej efektywna<sup>5</sup> w porównaniu z drugą pod względem efektywności — metodą najdalszego sąsiedztwa (Malina, 2004). Jednak jeżeli metoda Warda zostanie zastosowana dla innej macierzy odległości niż macierz kwadratów odległości euklidesowych, to może wówczas utracić swoje własności, a ponadto wyniki mogą różnić się od tych, które uzyskuje się dla kwadratów odległości euklidesowych. Z tego powodu w badaniu zastosowano metodę najdalszego sąsiedztwa (*complete linkage, furthest neighbour*).

Punktem wyjścia metod aglomeracyjnych (Nowak, 1990; Panek, 2009) jest założenie, że każdy obiekt  $O_i \in \{O_1, \dots, O_n\}$  traktuje się jako jednoelementową grupę  $G_i, i \in \{1, \dots, n\}$ . Następnie w macierzy odległości  $\mathbf{D}$  poszukuje się pomiędzy wszystkimi parami obiektów odległości minimalnej  $d_{pq} = \min_{i,j} d_{ij}$  i łączy

te obiekty w jedną grupę dwuelementową  $G_r$ , co powoduje zmniejszenie wyjściowej liczby grup o jeden. W kolejnym kroku wyznacza się odległości nowo utworzonej grupy  $G_r$  od wszystkich pozostałych grup. Odległości te wstawia się do macierzy  $\mathbf{D}$  w miejsce  $p$ -tego wiersza i  $p$ -tej kolumny, eliminując równocześnie wiersz i kolumnę o numerze  $q$ . Powstaje nowa, aktualna na danym etapie macierz odległości o coraz mniejszym wymiarze. Procedurę powtarza się do momentu, gdy wszystkie obiekty utworzą jedną grupę. Ogólna formuła wyznaczania odległości nowo utworzonej grupy  $G_r$ , powstałej z połączenia grup  $G_p$  i  $G_q$ , od pozostałych grup jest następująca:

<sup>5</sup> Przez efektywność metod taksonomicznych należy rozumieć zdolność prawidłowego rozpoznawania rzeczywistej struktury obiektów w wielowymiarowych przestrzeniach cech, którą bada się na przykładach empirycznych lub na podstawie zbiorów sztucznie wygenerowanych (Malina, 2004).

$$d_{ir} = a_i d_{ip} + a_q d_{iq} + b d_{pq} + c |d_{ip} - d_{iq}| \quad (15)$$

gdzie:

$d_{ir}, d_{ip}, d_{iq}$  — odległości między grupami  $G_i, G_r; G_i, G_p$  oraz  $G_i, G_q$ ,  
 $a_i, a_q, b, c$ , — parametry przekształcenia, charakterystyczne dla różnych metod tworzenia skupień.

W przypadku metody najdalszego sąsiedztwa wzór (15) przyjmie postać:

$$d_{ir} = \frac{1}{2} d_{ip} + \frac{1}{2} d_{iq} + \frac{1}{2} |d_{ip} - d_{iq}| \quad (16)$$

Metoda najdalszego sąsiedztwa skupia się na wewnętrznej spójności grup oraz wykazuje tendencję do znajdowania grup spójnych, tworząc czasami kilka równoległych małych skupień. Jest ona jednak czuła na występowanie jednostek oddalonych.

#### *OCENA GRUPOWANIA NA PODSTAWIE DENDROGRAMU ZA POMOCĄ KORELACJI KOFENETYCZNEJ*

Zastosowanie do macierzy odległości **D** jednej z metod aglomeracyjnych daje w wyniku dendrogram, który ilustruje kolejność łączenia się obiektów w klasy i poziomy, na których obiekty po raz pierwszy się połączyły. Wiele informacji z wyjściowych macierzy odległości między obiektami jest traconych w wyniku przeliczania odległości i tworzenia dendrogramu, zatem można zadać pytanie, czy dendrogram otrzymany za pomocą danej metody grupowania hierarchicznego dobrze pasuje do macierzy odległości, tzn. jak dobrze wyjściowa macierz odległości jest utrwalona na dendrogramie (Balicki, 2009).

W literaturze przedmiotu można znaleźć liczne propozycje dotyczące metod mierzenia dopasowania dendrogramu wynikającego z zastosowania określonej procedury grupowania hierarchicznego do macierzy odległości. Należą do nich np.: współczynnik korelacji kofenetycznej, wskaźnik całościowego dopasowania, statystyka Gowera, miara STRESS, współczynnik gamma Goodmana-Kruskala (Everitt, Dunn, 1991; Gordon, 1991). Miary te oparte są na porównaniu wyjściowej macierzy odległości i macierzy odległości na dendrogramie. Na podstawie dendrogramu można zapisać macierz odległości poziomów łączenia  $C = [c_{rs}]$ ,  $rs \in \{1, \dots, n\}$ , zwaną macierzą dendrogramu lub też macierzą kofenetyczną, której elementami są odległości odpowiadające poziomom łączenia, na których pary obiektów łączą się w tym samym skupieniu pierwszy raz. Spełniają one nierówność ultrametryczną (Balicki, 2009):

$$c_{rs} \leq \max(c_{rw}, c_{sw}) \quad r, s, w \in \{1, \dots, n\} \quad (17)$$

Dopasowanie można określić za pomocą współczynnika korelacji według momentu iloczynowego między odpowiadającymi sobie odległościami w obu macierzach **D** i **C**, który nosi nazwę współczynnika korelacji kofenetycznej:

$$r_{kof.} = \frac{\frac{n(n-1)}{2} \sum_{\substack{r,s \\ r < s}} d_{rs} \cdot c_{rs} - \sum_{\substack{r,s \\ r < s}} d_{rs} \cdot \sum_{\substack{r,s \\ r < s}} c_{rs}}{\sqrt{\left[ \frac{n(n-1)}{2} \sum_{\substack{r,s \\ r < s}} d_{rs}^2 - \left( \sum_{\substack{r,s \\ r < s}} d_{rs} \right)^2 \right] \left[ \frac{n(n-1)}{2} \sum_{\substack{r,s \\ r < s}} c_{rs}^2 - \left( \sum_{\substack{r,s \\ r < s}} c_{rs} \right)^2 \right]}} \quad (18)$$

gdzie sumowanie dotyczy wszystkich  $\frac{n(n-1)}{2}$  par obiektów.

Współczynnik ten można zapisać w postaci:

$$r_{kof.} = \frac{\sum_{\substack{r,s \\ r < s}} (d_{rs} - \bar{d})(c_{rs} - \bar{c})}{\sqrt{\sum_{\substack{r,s \\ r < s}} (d_{rs} - \bar{d})^2 \cdot \sum_{\substack{r,s \\ r < s}} (c_{rs} - \bar{c})^2}} \quad (19)$$

gdzie:

$$\bar{d} = \left( \frac{n(n-1)}{2} \right)^{-1} \cdot \sum_{\substack{r,s \\ r < s}} d_{rs} \quad \bar{c} = \left( \frac{n(n-1)}{2} \right)^{-1} \cdot \sum_{\substack{r,s \\ r < s}} c_{rs}$$

Wysokie wartości współczynnika wskazują, że dendrogram daje dobre podsumowanie różnic między obiektami lub podobieństw obiektów. Wartość współczynnika przeciętnie maleje wraz ze wzrostem liczby obiektów i prawie nie zależy od liczby zmiennych.

#### PRZYKŁAD EMPIRYCZNY

Przykładem zilustrowano opisane metody pod względem średniego poziomu rozwoju społeczno-gospodarczego w latach 2002—2009 obserwowanego w województwach. W badaniu wykorzystano dane statystyczne udostępnione przez GUS w Banku Danych Lokalnych. Do analizy przyjęto następujący zestaw cech diagnostycznych:

- $X_1$  — liczba ludności na 1 km<sup>2</sup>,
- $X_2$  — udział ludności w wieku przedprodukcyjnym w ogólnej liczbie ludności,
- $X_3$  — udział ludności w wieku produkcyjnym w ogólnej liczbie ludności,
- $X_4$  — udział ludności w wieku poprodukcyjnym w ogólnej liczbie ludności,
- $X_5$  — ludność w wieku nieprodukcyjnym na 100 osób w wieku produkcyjnym,

- $X_6$  — ludność w wieku poprodukcyjnym na 100 osób w wieku przedprodukcyjnym,
- $X_7$  — ludność w wieku poprodukcyjnym na 100 osób w wieku produkcyjnym,
- $X_8$  — liczba kobiet na 100 mężczyzn,
- $X_9$  — zgony na 1000 ludności,
- $X_{10}$  — przyrost naturalny na 1000 ludności,
- $X_{11}$  — urodzenia żywe na 1000 ludności,
- $X_{12}$  — liczba małżeństw zawartych w ciągu roku na 1000 ludności,
- $X_{13}$  — liczba rozwodów na 1000 ludności,
- $X_{14}$  — zgony niemowląt na 1000 urodzeń żywych,
- $X_{15}$  — przeciętne miesięczne wydatki na osobę,
- $X_{16}$  — stopa bezrobocia w %,
- $X_{17}$  — liczba ofert pracy ogółem na bezrobotnego,
- $X_{18}$  — wskaźnik zatrudnienia ogółem w %,
- $X_{19}$  — przeciętne miesięczne wynagrodzenie brutto w relacji do średniej krajowej (Polska=100),
- $X_{20}$  — przeciętna powierzchnia użytkowa mieszkania w m<sup>2</sup> na osobę,
- $X_{21}$  — liczba mieszkań na 10 tys. ludności,
- $X_{22}$  — liczba studentów na 10 tys. ludności,
- $X_{23}$  — liczba praktyk lekarskich w miastach na 10 tys. ludności,
- $X_{24}$  — liczba praktyk lekarskich na wsi na 10 tys. ludności,
- $X_{25}$  — liczba ludności na 1 aptekę ogólnodostępną,
- $X_{26}$  — liczba osób korzystających ze świadczeń pomocy społecznej na 10 tys. ludności,
- $X_{27}$  — stopień wykorzystania miejsc noclegowych w %,
- $X_{28}$  — czytelnicy bibliotek publicznych na 1000 ludności,
- $X_{29}$  — wypożyczenia księgozbioru na czytelnika,
- $X_{30}$  — liczba ludności na 1 miejsce w kinach stałych,
- $X_{31}$  — liczba widzów i słuchaczy w teatrach i instytucjach muzycznych na 1000 ludności,
- $X_{32}$  — PKB na mieszkańca,
- $X_{33}$  — nakłady inwestycyjne na mieszkańca w zł,
- $X_{34}$  — drogi publiczne o twardej nawierzchni w km na 100 km<sup>2</sup> powierzchni,
- $X_{35}$  — liczba samochodów osobowych zarejestrowanych na 1000 ludności,
- $X_{36}$  — liczba ofiar śmiertelnych na 100 wypadków drogowych,
- $X_{37}$  — emisja przemysłowych zanieczyszczeń pyłowych powietrza w tonach na 100 km<sup>2</sup>,
- $X_{38}$  — emisja przemysłowych zanieczyszczeń gazowych powietrza w tonach na 100 km<sup>2</sup>,
- $X_{39}$  — udział parków narodowych w ogólnej powierzchni w %,
- $X_{40}$  — plony z 1 ha zbóż ogółem w dt,
- $X_{41}$  — dochody budżetu województwa ogółem na mieszkańca w zł,
- $X_{42}$  — wydatki z budżetu województwa ogółem na mieszkańca w zł,
- $X_{43}$  — liczba podmiotów gospodarczych na 1000 mieszkańców.

W pierwszym kroku wyznaczono wartości współczynników zmienności dla każdej cechy w kolejnych latach oraz średnie wartości współczynników zmienności według wzoru (5). Następnie wyeliminowano cechy  $X_8, X_{19}, X_{20}$ , których współczynniki zmienności oscylowały w pobliżu wartości średniej współczynnika zmienności w sposób przypadkowy, co oznacza, że były one prawie stałe w czasie. W kolejnym kroku dokonano estymacji trendów o postaci liniowej, zgodnie ze wzorem (7) oraz zbadano istotność współczynników kierunkowych. Usunięto cechy, dla których współczynniki kierunkowe trendu liniowego były nieistotne:  $X_9, X_{14}, X_{16}, X_{17}, X_{24}, X_{31}, X_{37}, X_{38}, X_{40}, X_{42}$ . Dla pozostałych cech wyznaczono prognozy współczynników zmienności i wyeliminowano te, dla których prognozowane współczynniki zmienności za pomocą trendu liniowego były mniejsze od 20%.

Po tym etapie pozostały następujące cechy:  $X_1, X_{21}, X_{22}, X_{23}, X_{26}, X_{30}, X_{32}, X_{33}, X_{34}, X_{36}, X_{39}$ .

Oprócz badania zmienności cech w kolejnych latach przeprowadzono analizę powiązań między zmiennymi w ujęciu dynamicznym. Wyznaczono prognozy wartości współczynników korelacji na podstawie trendów liniowych, w których zaobserwowano istotne współczynniki kierunkowe. Na tej podstawie zbudowano macierz prognoz współczynników korelacji, którą przedstawiono w tabl. 1.

**TABL. 1. MACIERZ PROGNOZ WSPÓŁCZYNNIKÓW KORELACJI POMIĘDZY CECAMI**

|          | $X_1$  | $X_{21}$ | $X_{22}$ | $X_{23}$ | $X_{26}$ | $X_{30}$ | $X_{32}$ | $X_{33}$ | $X_{34}$ | $X_{36}$ | $X_{39}$ |
|----------|--------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| $X_1$    | 1,000  | -0,102   | 0,237    | -0,471   | -0,520   | -0,682   | 0,326    | 0,335    | 0,874    | -0,671   | -0,266   |
| $X_{21}$ | -0,102 | 1,000    | 0,600    | -0,369   | -0,087   | -0,542   | 0,640    | 0,683    | -0,284   | -0,011   | 0,228    |
| $X_{22}$ | 0,237  | 0,600    | 1,000    | -0,775   | -0,617   | -0,647   | 0,556    | 0,569    | 0,362    | -0,372   | 0,029    |
| $X_{23}$ | -0,471 | -0,369   | -0,775   | 1,000    | 0,450    | 0,451    | -0,587   | -0,534   | -0,621   | 0,404    | -0,035   |
| $X_{26}$ | -0,520 | -0,087   | -0,617   | 0,450    | 1,000    | 0,629    | -0,519   | -0,446   | -0,637   | 0,427    | 0,090    |
| $X_{30}$ | -0,682 | -0,542   | -0,647   | 0,451    | 0,629    | 1,000    | -0,720   | -0,670   | -0,472   | 0,509    | 0,166    |
| $X_{32}$ | 0,326  | 0,640    | 0,556    | -0,587   | -0,519   | -0,720   | 1,000    | 0,908    | 0,251    | -0,164   | -0,310   |
| $X_{33}$ | 0,335  | 0,683    | 0,569    | -0,534   | -0,446   | -0,670   | 0,908    | 1,000    | 0,182    | -0,336   | -0,275   |
| $X_{34}$ | 0,874  | -0,284   | 0,362    | -0,621   | -0,637   | -0,472   | 0,251    | 0,182    | 1,000    | -0,607   | -0,180   |
| $X_{36}$ | -0,671 | -0,011   | -0,372   | 0,404    | 0,427    | 0,509    | -0,164   | -0,336   | -0,607   | 1,000    | 0,204    |
| $X_{39}$ | -0,266 | 0,228    | 0,029    | -0,035   | 0,090    | 0,166    | -0,310   | -0,275   | -0,180   | 0,204    | 1,000    |

Źródło: obliczenia własne.

Wyznaczono macierz odwrotną (tabl. 2) do macierzy prognoz współczynników korelacji pomiędzy cechami i wyłączono z badania te cechy, dla których elementy diagonalne macierzy odwrotnej są większe od 10, co jest symptomem złego uwarunkowania numerycznego macierzy  $\mathbf{R}$ :  $X_{21}, X_{23}, X_{30}, X_{34}$ .

**TABL. 2. MACIERZ ODWROTNA DO MACIERZY PROGNOZ WSPÓŁCZYNNIKÓW KORELACJI POMIĘDZY CECAMI**

|          | $X_1$   | $X_{21}$ | $X_{22}$ | $X_{23}$ | $X_{26}$ | $X_{30}$ | $X_{32}$ | $X_{33}$ | $X_{34}$ | $X_{36}$ | $X_{39}$ |
|----------|---------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| $X_1$    | 8,896   | 2,901    | -6,780   | -11,587  | -8,554   | -0,514   | -8,884   | -1,011   | -14,158  | 2,438    | -3,726   |
| $X_{21}$ | 2,901   | 83,626   | -12,427  | 27,851   | -16,588  | 48,942   | -5,695   | -21,767  | 57,020   | -0,251   | -20,965  |
| $X_{22}$ | -6,780  | -12,427  | 0,222    | -5,736   | 3,101    | -12,964  | -6,647   | 8,151    | -3,607   | 2,096    | 1,786    |
| $X_{23}$ | -11,587 | 27,851   | -5,736   | 13,753   | -1,569   | 9,435    | -7,185   | 0,136    | 33,448   | 1,218    | -6,612   |
| $X_{26}$ | -8,554  | -16,588  | 3,101    | -1,569   | 9,070    | -13,528  | 0,158    | 7,961    | -0,292   | 0,424    | 4,876    |
| $X_{30}$ | -0,514  | 48,942   | -12,964  | 9,435    | -13,528  | 25,605   | -10,454  | -9,338   | 30,753   | 0,954    | -14,058  |
| $X_{32}$ | -8,884  | -5,695   | -6,647   | -7,185   | 0,158    | -10,454  | 0,300    | -0,446   | -1,532   | -1,483   | 0,586    |
| $X_{33}$ | -1,011  | -21,767  | 8,151    | 0,136    | 7,961    | -9,338   | -0,446   | 13,245   | -7,382   | 1,883    | 7,072    |
| $X_{34}$ | -14,158 | 57,020   | -3,607   | 33,448   | -0,292   | 30,753   | -1,532   | -7,382   | 65,843   | 0,242    | -11,232  |
| $X_{36}$ | 2,438   | -0,251   | 2,096    | 1,218    | 0,424    | 0,954    | -1,483   | 1,883    | 0,242    | 2,786    | 0,026    |
| $X_{39}$ | -3,726  | -20,965  | 1,786    | -6,612   | 4,876    | -14,058  | 0,586    | 7,072    | -11,232  | 0,026    | 6,482    |

Źródło: jak przy tabl. 1.

Po eliminacji otrzymano nową zredukowaną macierz prognoz współczynników korelacji oraz macierz do niej odwrotną, w której wszystkie elementy diagonalne były mniejsze od 3. W wyniku przeprowadzonego postępowania otrzymano nową listę cech:  $X_1$ ,  $X_{22}$ ,  $X_{26}$ ,  $X_{32}$ ,  $X_{36}$ ,  $X_{39}$ , które są słabo skorelowane między sobą i równocześnie silnie skorelowane z pozostałymi cechami, niewybranymi jako cechy diagnostyczne.

Uwzględniając cel badania, w kolejnym kroku wyznaczono dla lat badanego okresu macierz odległości Penrose'a **D** (tabl. 3) zgodnie ze wzorem (13).

**TABL. 3. MACIERZ ODLEGŁOŚCI PENROSE'A**

|      | 2002   | 2003   | 2004   | 2005   | 2006   | 2007   | 2008   | 2009   |
|------|--------|--------|--------|--------|--------|--------|--------|--------|
| 2002 | 0,0000 | 0,0127 | 0,0730 | 0,1246 | 0,2694 | 0,4213 | 0,7280 | 0,8620 |
| 2003 | 0,0127 | 0,0000 | 0,0474 | 0,0672 | 0,1803 | 0,1126 | 0,6473 | 0,5231 |
| 2004 | 0,0730 | 0,0474 | 0,0000 | 0,0353 | 0,0353 | 0,1902 | 0,4011 | 0,5231 |
| 2005 | 0,1246 | 0,0672 | 0,0353 | 0,0000 | 0,0578 | 0,1371 | 0,3836 | 0,5048 |
| 2006 | 0,2694 | 0,1803 | 0,0353 | 0,0578 | 0,0000 | 0,1505 | 0,4279 | 0,5282 |
| 2007 | 0,4213 | 0,1126 | 0,1902 | 0,1371 | 0,1505 | 0,0000 | 0,0725 | 0,1319 |
| 2008 | 0,7280 | 0,6473 | 0,4011 | 0,3836 | 0,4279 | 0,0725 | 0,0000 | 0,0189 |
| 2009 | 0,8620 | 0,5231 | 0,5231 | 0,5048 | 0,5282 | 0,1319 | 0,0189 | 0,0000 |

Źródło: jak przy tabl. 1.

Na podstawie macierzy odległości Penrose'a został utworzony dendrogram, przedstawiający hierarchiczne uporządkowanie lat (wykres). Analizując macierz odległości Penrose'a i przyjmując za podstawę porównań rok 2002, można stwierdzić, że zmiany w poziomie rozwoju społeczno-gospodarczego województw w 2003 r. były co najmniej pięć razy niższe niż w kolejnych latach.

Za pomocą współczynnika korelacji kofenetycznej (wzór 18), wyznaczonego poprzez porównanie odległości w wyjściowej macierzy odległości oraz macierzy kofenetycznej (tabl. 4), oceniono stopień dopasowania dendrogramu do macierzy odległości, tzn. zbadano, czy jest ona dobrze odzwierciedlona na dendrogramie. Wartość współczynnika korelacji kofenetycznej wyniosła 0,74, co świadczy o tym, iż dendrogram daje dobre podsumowanie obserwowanych różnic między latami. Zatem otrzymany dendrogram może posłużyć do grupowania lat.

**TABL. 4. MACIERZ KOFENETYCZNA (dendrogramu)**

|      | 2002  | 2003  | 2004  | 2005  | 2006  | 2007  | 2008  | 2009  |
|------|-------|-------|-------|-------|-------|-------|-------|-------|
| 2002 | 0,000 | 0,012 | 0,269 | 0,269 | 0,269 | 0,862 | 0,862 | 0,862 |
| 2003 | 0,012 | 0,000 | 0,269 | 0,269 | 0,269 | 0,862 | 0,862 | 0,862 |
| 2004 | 0,269 | 0,269 | 0,000 | 0,035 | 0,058 | 0,862 | 0,862 | 0,862 |
| 2005 | 0,269 | 0,269 | 0,035 | 0,000 | 0,058 | 0,862 | 0,862 | 0,862 |
| 2006 | 0,269 | 0,269 | 0,058 | 0,058 | 0,000 | 0,862 | 0,862 | 0,862 |
| 2007 | 0,862 | 0,862 | 0,862 | 0,862 | 0,862 | 0,000 | 0,132 | 0,132 |
| 2008 | 0,862 | 0,862 | 0,862 | 0,862 | 0,862 | 0,132 | 0,000 | 0,019 |
| 2009 | 0,862 | 0,862 | 0,862 | 0,862 | 0,862 | 0,132 | 0,019 | 0,000 |

Źródło: jak przy tabl. 1.

Liczbę klas określono za pomocą badania ilorazów sąsiednich wartości w uporządkowanym nierosnąco ciągu odległości pomiędzy grupami (Grabiński, 1992) otrzymanymi w wyniku zastosowania metody najdalszego sąsiedztwa. Otrzymano cztery skupienia lat: {2002, 2003}, {2004, 2005, 2006}, {2007}, {2008, 2009}.

Analizując dendrogram można zauważyć, że najmniejsza odległość wystąpiła pomiędzy latami 2002 i 2003. W latach tych zaobserwowano najniższą liczbę studentów przypadających na 10 tys. ludności, najniższą wartość PKB *per capita* oraz najwyższy udział powierzchni parków narodowych w ogólnej powierzchni województw. W latach 2004—2006 z kolei odnotowano najwyższą liczbę ofiar śmiertelnych na 100 wypadków drogowych, najwyższą liczbę studentów na 10 tys. ludności oraz najwyższą liczbę osób korzystających ze świadczeń pomocy społecznej na 10 tys. ludności. W ostatniej grupie znalazły się lata 2008 i 2009, w których charakterystyczną cechą jest najwyższy PKB *per capita* oraz najniższa liczba osób korzystających ze świadczeń pomocy społecznej na 10 tys. mieszkańców. Rok 2007 tworzy odrębną grupę, ponieważ wtedy wartość poszczególnych cech jest zbliżona do wartości średnich ogólnych dla wszystkich lat.

Analiza tendencji rozwojowych poszczególnych cech pozwoliła stwierdzić, że z roku na rok następował wzrost wartości PKB *per capita* oraz zdecydowany spadek liczby osób korzystających ze świadczeń pomocy społecznej na 10 tys. ludności. W ostatnich dwóch latach zaś zanotowano spadek liczby ofiar śmiertelnych na 100 wypadków drogowych.

## Podsumowanie

W artykule przedstawiono dynamiczne ujęcie doboru cech diagnostycznych w klasyfikacji lat pod względem średniego poziomu rozwoju społeczno-gospodarczego w województwach Polski. Klasyfikację przeprowadzono oceniając podobieństwo poszczególnych lat za pomocą odległości Penrose'a. Dopasowanie dendrogramu, otrzymanego za pomocą metody najdalszego sąsiedztwa, zmierzono wykorzystując współczynnik korelacji kofenetycznej, którego wartość wyniosła 0,74. Świadczy to, iż dendrogram daje dobre podsumowanie obserwowanych różnic między latami i może być podstawą grupowania lat. Analizując dendrogram otrzymano cztery skupienia lat. Wśród otrzymanych klas na uwagę zasługuje skupienie obejmujące lata 2008 i 2009, w których charakterystyczną cechą jest najwyższy PKB *per capita* oraz najniższa liczba osób korzystających ze świadczeń pomocy społecznej na 10 tys. mieszkańców, co może sugerować, że w tym okresie średni poziom rozwoju społeczno-gospodarczego województw był, w porównaniu z pozostałymi latami, na najwyższym poziomie.

## LITERATURA

- Balicki A. (2009), *Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne*, Wydawnictwo Uniwersytetu Gdańskiego
- Everitt B. S., Dunn G. (1991), *Applied Multivariate Data Analysis*, Edward Arnold, London
- Gordon A. D. (1991), *Classification*, Chapman & Hall/CRC, Boca Raton
- Grabiński T. (1992), *Metody taksonometrii*, Akademia Ekonomiczna w Krakowie
- Grabiński T., Wydymus S., Zeliaś A. (1989), *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*, Państwowe Wydawnictwo Naukowe, Warszawa
- Malina A. (2004), *Wielowymiarowa analiza przestrzennego zróżnicowania struktury gospodarki Polski według województw*, Wydawnictwo Akademii Ekonomicznej w Krakowie
- Malina A., Zeliaś A. (1997), *O budowie taksonomicznej miary jakości życia*, [w:] *Taksonomia 4*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu
- Młodak A. (2006), *Analiza taksonomiczna w statystyce regionalnej*, DIFIN, Warszawa
- Nowak E. (1990), *Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych*, Państwowe Wydawnictwo Ekonomiczne, Warszawa
- Ostasiewicz W. (red.) (1998), *Statystyczne metody analizy danych*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu
- Panek T. (2009), *Statystyczne metody wielowymiarowej analizy porównawczej*, SGH
- Strahl D. (red.) (2006), *Metody oceny rozwoju regionalnego*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu
- Ward J. H. (1963), *Hierarchical grouping to optimize an objective function*, „Journal of the American Statistical Association”, No. 58
- Zeliaś A. (red.) (2000), *Taksonomiczna analiza przestrzennego zróżnicowania poziomu życia w Polsce w ujęciu dynamicznym*, Wydawnictwo Akademii Ekonomicznej w Krakowie

## SUMMARY

*The primary objective of the article is to present results of taxonomic studies, which use a dynamic approach in the selection of features. The resulting set of diagnostic features was used for classification in years 2002–2009 the average level of socio-economic development observed in the Polish voivodships. Procedure was applied taking into account changes in the relationship between adopted variables.*

*The result is a set of diagnostic features, used to forecast the development of the phenomenon under investigation. The classification was carried out by assessing the similarity of years in terms of the average level of socio-economic development in the regions with the Penrose distance. Adjusting the dendrogram (tree diagram), obtained by application of the furthest neighborhood method, was measured using the cophenetic correlation coefficient.*

## РЕЗЮМЕ

*Основной целью статьи является представление результатов таксономического обследования, в котором был использован динамический подход к подбору характеристик. Полученный набор диагностических характеристик был использован в классификации 2002—2009 гг. в отношении к среднему уровню социально-экономического развития наблюдаемого в воеводствах Польши.*

*Была использована процедура учитывающая изменения в отношениях между принятыми переменными, в результате которой получается набор диагностических характеристик, который служит прогнозированию развития обследуемого явления. Классификация была проведена с учетом сходства лет в отношении к среднему уровню социально-экономического развития в воеводствах с помощью расстояния Penrose'a. Соответствие дендрограммы, полученной в результате использования метода дальних районов (окрестностей), было измерено с использованием коэффициента кофенетической корреляции.*