PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 755 — 1997
Gospodarka lokalna w teorii i praktyce

Elżbieta Sobczak
Akademia Ekonomiczna we Wrocławiu

PROPOZYCJA KOMPLEKSOWEJ ANALIZY PORÓWNAWCZEJ POZIOMU ZJAWISKA EKONOMICZNEGO WYBRANYCH OBIEKTÓW

1. Wstęp

Zmiany poziomu rozwoju pewnych zjawisk ekonomicznych mogą wiązać się ze zmianami strukturalnymi, a także z rozwojem gospodarczym, dlatego stanowią dziedzinę zainteresowania polityki gospodarczej.

Metody wielowymiarowej analizy porównawczej umożliwiają wyodrębnienie regionów o podobnym poziomie rozwoju analizowanego zjawiska ekonomicznego i łączenie ich w homogeniczne grupy, a także wyodrębnienie faz rozwoju zjawiska ekonomicznego w badanych regionach.

Zakłada się, że analizie podlegać będzie n obiektów-regionów, ze względu na poziom pewnego zjawiska ekonomicznego. Pomiaru dokonano dla t okresów badania.

Niech obrazem liczbowym badanego zjawiska ekonomicznego opisanego za pomocą jednej cechy statystycznej realizowanej w t okresach, będzie następująca macierz:

$$Y_{..} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1t} \\ y_{21} & y_{22} & \cdots & y_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nt} \end{bmatrix}_{(n \times t)} \tag{1}$$

gdzie: y_{rl} – poziom badanego zjawiska w r-tym obiekcie i l-tym okresie badania;
$r, s = 1, \ldots, n$ (numer obiektu badania);
$l, l = 1, \ldots, t$ (numer okresu badania).

Klasyfikacja przestrzenna obejmuje podział zbioru obiektów-regionów na klasy, ze względu na cechę statystyczną b_1, zrealizowaną w l-tym okresie. Analizie poddana zostanie l-ta kolumna macierzy (1).

Klasyfikacja przestrzenno-czasowa to podział obiektów-regionów na klasy ze względu na poziom cechy b_1, ustalonej w t badanych okresach. Punktem wyjścia analizy porównawczej jest w tym przypadku macierz (1).

Zgodnie z przyjętą metodologią, klasyfikacja czasowa została potraktowana jako podział okresu badania na fazy rozwoju, czyli podokresy kolejno po sobie następujących momentów czasu, podobnych ze względu na poziom badanej cechy statystycznej b_1, którą jest poziom zjawiska ekonomicznego.

2. Metodologia przestrzennej i przestrzenno-czasowej klasyfikacji obiektów-regionów ze względu na poziom analizowanego zjawiska ekonomicznego

Do przeprowadzenia klasyfikacji obiektów-regionów ze względu na poziom wybranego zjawiska ekonomicznego w układzie przestrzennym i przestrzenno-czasowym, wykorzystano schemat postępowania składający się z następujących kroków:

1) Normalizacja cechy statystycznej odnoszącej się do wszystkich badanych okresów

W tym celu wykorzystano następującą formułę:

$$p_{rl} = \frac{y_{rl}}{\sum_{r=1}^{n} y_{rl}}, \tag{2}$$

gdzie: $r = 1, \ldots, n$ (numer obiektu badania),
$l = 1, \ldots, t$ (numer okresu badania),
y_{rl} – wartość liczbowa cechy statystycznej b_1 w r-tym obiekcie, w l-tym okresie badania,
p_{rl} – unormowana wartość liczbowa cechy statystycznej b_1 w r-tym obiekcie, w l-tym okresie badania.

W taki sposób powstała poniższa macierz unormowanych danych statystycznych:

$$P.. = \begin{bmatrix} p_{11} & p_{12} & \cdots & y_{1t} \\ p_{21} & p_{22} & \cdots & p_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nt} \end{bmatrix}_{(n \times t)} \tag{3}$$

gdzie: p_{rl} – unormowana wartość liczbowa cechy statystycznej b_1 w r-tym obiekcie, w l-tym okresie badania.

2) Wybór miary odległości

Do obliczenia odległości między badanymi obiektami wykorzystana została miara Braya i Curtisa [2]. Odległości będące podstawą klasyfikacji badanych obiektów (regionów) w układzie przestrzennym obliczono zgodnie z poniższym wzorem:

$$d'_{rs} = \frac{|p_{rl} - p_{sl}|}{p_{rl} + p_{sl}}, \tag{4}$$

gdzie: $r, s = 1, \ldots, n$ (numer obiektu badania),
$l = 1, \ldots, t$ (numer okresu badania),
d'_{rs} — odległość między r-tym i s-tym obiektem badania ze względu na poziom cechy b_1, w l-tym okresie,
p_{rl} — unormowana wartość liczbowa cechy statystycznej b_1 w r-tym obiekcie w l-tym okresie badania.

Odległości będące podstawą klasyfikacji badanych obiektów w układzie przestrzenno-czasowym obliczono według następującej formuły:

$$D'_{rs} = D(p_{r\cdot}, p_{s\cdot}) = \frac{\sum_{l=1}^{t} |p_{rl} - p_{sl}|}{\sum_{l=1}^{t} (p_{rl} + p_{sl})}, \tag{5}$$

gdzie: $r, s = 1, \ldots, n$ (numer obiektu badania),
$l = 1, \ldots, t$ (numer okresu badania),
d'_{rs} — odległość między r-tym i s-tym obiektem badania ze względu na poziom cechy b_1, w t okresach,
$p_{r\cdot}, p_{s\cdot}$ — odpowiednio r-ty i s-ty wektor wierszowy macierzy unormowanych danych statystycznych (por. macierz (3)),
p_{rl} — unormowana wartość liczbowa cechy statystycznej b_1 w r-tym obiekcie, l-tym okresie badania.

W ten sposób dla każdego okresu badania powstaje macierz odległości o wymiarach $(n \times n)$, składająca się z elementów obliczonych według wzoru (4), będąca podstawą klasyfikacji przestrzennej.

Natomiast macierz odległości o wymiarach $(n \times n)$, której elementy oblicza się zgodnie z formułą (5) jest podstawą klasyfikacji przestrzenno-czasowej regionów.

3) Wybór metody klasyfikacji

Klasyfikacji przestrzennej i przestrzenno-czasowej regionów ze względu na poziom zjawiska ekonomicznego można dokonać posługując się metodą najdalszego sąsiada. Zdecydowano się na wybór metody hierarchicznej, ponieważ w wyniku jej zastosowania otrzymuje się ciąg klasyfikacji, od podziału w którym każdy obiekt stanowi odrębną klasę, aż do podziału w którym wszystkie obiekty

znajdują się w jednej klasie. Umożliwia to kontrolę procesu klasyfikacji i wybór optymalnego jej etapu.

Przyjęto, że podział na klasy uznaje się za przydatny jeżeli spełnia następujące warunki:

– liczba klas jednoelementowych nie jest zbyt duża;
– nie wystepują klasy zbyt liczne.

Zastosowanie hierarchicznej metody klasyfikacji obiektów stwarza konieczność ustalenia tzw. reguły stop do wyboru z uzyskanego ciągu podziałów, podziału optymalnego. Proponuje się tutaj zastosowanie poniższego wskaźnika jakości klasyfikacji J.J. Fortiera i H. Solomona [3]:

$$J_e = \sum_{Ge_u} \sum_{\substack{\omega_r, \omega_s \in G_u^e \\ r<s}} d_{rs} - \frac{d^*}{2} \sum_{G_u^e} n_u^e (n_u^e - 1), \tag{6}$$

gdzie: J_e – wskaźnik jakości klasyfikacji dla e-tego etapu procedury hierarchicznej,
$u = 1, \ldots, U$ (numer klasy),
$r, s = 1, \ldots, n$ (numer obiektu badania),
$\Omega = \{\omega_1, \omega_2, \ldots, \omega_r, \ldots, \omega_n\}$ (zbiór obiektów badania),
G_u^e – u-ta klasa w e-tym etapie procedury hierarchicznej,
d_{rs} – odległość między r-tym i s-tym obiektem ze względu na poziom zjawiska ekonomicznego uzyskany w l-tym roku,
d^* – krytyczna wartość miary odległości,
n_u^e – liczebność klasy G_u w e-tym kroku procedury hierarchicznej.

Przy klasyfikacji przestrzenno-czasowej regionów formuła (6) ulega modyfikacji, w miejsce miary d_{rs} podstawia się D_{rs} – odległość między r-tym i s-tym obiektem, ze względu na poziom zjawiska ekonomicznego uzyskany w t okresach badania.

W ciągu podziałów ten jest najlepszy, dla którego wskaźnik jakości klasyfikacji przyjmuje najmniejszą wartość.

Do ustalenia podziału optymalnego z ciągu klasyfikacji, zgodnie z propozycją J.J. Fortiera i H. Solomona, niezbędne było przyjęcie tzw. odległości krytycznej. Sposób jej obliczenia prezentuje poniższa formuła:

$$d^* = \frac{1}{t+1} \left(\sum_{l=1}^{t} \max_s \min_r [d_{rs}]_l + \max_s \min_r [D_{rs}] \right), \tag{7}$$

gdzie: d^* – odległość krytyczna,
t – liczba okresów badania,
$l = 1, \ldots, t$ (numer obiektu badania),
$r, s = 1, \ldots, n$ (numer obiektu badania),
$[d_{rs}]_l$) – macierz odległości między obiektami, ze względu na wartości cechy b_1 w l-tym okresie badania, będąca podstawą klasyfikacji przestrzennej,
$[D_{rs}]$ – macierz odległości między obiektami, ze względu na wartości cechy b_1 w t okresach badania, będąca podstawą klasyfikacji przestrzenno-czasowej.

4) Rejestracja zmian w czasie w klasyfikacjach przestrzennych regionów

Porównania wyników klasyfikacji uzyskanych metodą najdalszego sąsiada, na podstawie informacji statystycznych pochodzących z dwóch różnych okresów czasu, można dokonać posługując się różnymi metodami. Metoda najdalszego sąsiada jest metodą hierarchiczną, dlatego efektem jej zastosowania jest ciąg wyników klasyfikacji, z którego po zastosowaniu określonej reguły wybiera się jeden optymalny podział. W takim przypadku do porównania klasyfikacji regionów można wykorzystać następujące podejścia:

1. Porównanie dwóch ciągów klasyfikacji uzyskanych na podstawie informacji statystycznych pochodzących z l-tego i l-tego okresu badania.
2. Porównanie dwóch optymalnych podziałów wybranych z ciągów klasyfikacji uzyskanych dla l-tego i l-tego okresu badania.

Zgodność dwóch ciągów klasyfikacji proponuje się ocenić stosując następującą miarę [5]:

$$f(B^l, B^{l'}) = 1 - \frac{\sum_{r=2}^{n} \sum_{s=1}^{r-1} |b^l_{rs} - b^{l'}_{rs}|}{\binom{n}{2}}, \qquad 8)$$

gdzie: B^l, $B^{l'}$ – przekształcone macierze odległości między obiektami ze względu na cechę b_1, ustalone dla l-tego i l-tego okresu badania,

$r, s = 1, \ldots, n$ (numer obiektu badania),

b^l_{rs}, $b^{l'}_{rs}$ – elementy macierzy B^l, $B^{l'}$ odpowiadające poziomom połączenia obiektów ω_r i ω_s w jedną klasę, odpowiednio w l-tym i l-tym okresie.

Miara (8) wykorzystuje tzw. przekształcone macierze odległości między obiektami ze względu na cechę b_1, ustalone dla l-tego i l-tego okresu. Każdy element przekształconej macierzy odległości informuje na jakim poziomie połączenia klas regiony znajdują się w jednej klasie.

Do konstrukcji macierzy odległości wykorzystano miarę braku podobieństwa Braya i Curtisa (por. formuła (4)), unormowaną w przedziale [0, 1]. Dlatego miara (8) może również przyjmować wartości z tego przedziału. Jeżeli miara zgodności ciągów klasyfikacji przyjmuje wartość równą 1 oznacza to, że ciągi klasyfikacji pochodzące z dwóch różnych okresów są identyczne. Im większe zachodzą między nimi różnice tym bardziej wartość miary zbliża się do zera.

Do oceny zgodności wyników klasyfikacji pod względem składu porównywanych klas proponuje się zastosowanie miary zgodności wyników klasyfikacji E. Nowaka (4), którą prezentuje poniższa formuła:

$$Z = \frac{1}{U + U'} \left(\sum_{u=1}^{U} \max_{u} \frac{n_{u,u'}}{\max\{n_u, n_{u'}\}} + \sum_{u'=1}^{U'} \max_{u} \frac{n_{u,u'}}{\max\{n_u, n_{u'}\}} \right), \qquad (9)$$

gdzie: $u = 1, \ldots, U$ numer klasy G_u w podziale pierwszym,
$u' = 1, \ldots, U'$ numer klasy $G_{u'}$ w podziale drugim,
n_u — liczebność klasy G_u w podziale pierwszym,
$n_{u'}$ — liczebność klasy $G_{u'}$ w podziale drugim
$n_{u,u'}$ — liczba obiektów należących do klasy G_u w jednym podziale i do klasy $G_{u'}$ w drugim z porównywanych podziałów.

Wskaźnik podobieństwa wyników podziałów E. Nowaka przyjmuje wartości z przedziału $\left[\frac{1}{n}, 1\right]$ (gdzie: n jest liczbą obiektów poddanych klasyfikacji). Miara osiąga najmniejszą wartość, gdy w jednym podziale wszystkie obiekty znajdują się w jednej klasie, a w drugim każdy obiekt stanowi odrębną klasę. Jednak nawet wówczas miara przyjmuje wartość , a nie zero. Uważa się bowiem, że występuje pewne podobieństwo między tymi podziałami. Klasa z podziału pierwszego posiada jeden element wspólny z każdą jednoelementową klasą podziału drugiego. Podobieństwo między tymi podziałami oceniane jest jako tym większe, im mniejsza liczba obiektów poddawana jest klasyfikacji. Ta własność miary zgodności wyników podziałów stanowi jej zaletę interpretacyjną. Im wyższą wartość przyjmuje miara, przy tej samej liczbie klasyfikowanych obiektów, tym większa jest zgodność wyników dwóch podziałów. Miara przyjmuje wartość 1, gdy porównywane podziały są identyczne.

3. Metodologia klasyfikacji czasowej (periodyzacji) obiektów-regionów

Punktem wyjścia periodyzacji poziomu zjawiska ekonomicznego jest r-ty wiersz macierzy danych wyjściowych (1). Do wyodrębnienia faz rozwoju poziomu zjawiska ekonomicznego zastosowano schemat postępowania składający się z następujących kroków:

1) Normalizacja danych wyjściowych
Do tego celu wykorzystano poniższą formułę:

$$p'_{rl} = \frac{y_{rl}}{\sum_{l=1}^{t} y_{rl}}, \tag{10}$$

gdzie: $r = 1, \ldots, n$ (numer obiektu badania),
$l = 1, \ldots, t$ (numer okresu badania),
y_{rl} — wartość liczbowa cechy statystycznej b_1, w r-tym obiekcie, l-tym okresie badania,
p'_{rl} — wartość liczbowa cechy statystycznej b_1, w r-tym obiekcie, l-tym okresie badania, po unormowaniu będącym podstawą periodyzacji.

W ten sposób dla każdego obiektu badania powstał następujący wektor unormowanych wartości cechy b_1:

$$p'_{r\cdot} = [p'_{r1}\, p'_{r2} \ldots p'_{rt}], \tag{11}$$

gdzie: $p'_{r\cdot}$ – wektor unormowanych wartości cechy b1, będący podstawą jej periodyzacji w r-tym obiekcie.

2) Obliczenie odległości między wartościami cechy b_1, odpowiadającymi poszczególnym okresom w danym obiekcie badania.

Do tego celu wykorzystano miarę odległości Braya i Curtisa w następującej zmodyfikowanej postaci:

$$d'_{ltr} = \frac{|p'_{rl} - p'_{rt}|}{p'_{rl} + p'_{rt}}, \tag{12}$$

gdzie: $r = 1, \ldots, n$ (numer obiektu badania),
$l, t = 1, \ldots, t$ (numer okresu badania),
d'_{ltr} – odległość między wartością cechy statystycznej b_1 r-tego obiektu, ustaloną dla l-tego i t-tego okresu badania,
p'_{rl}, p'_{rt} – wartość cechy statystycznej b_1 r-tego obiektu, odpowiednio w l-tym i t-tym okresie badania po unormowaniu będącym podstawą periodyzacji.

Dla każdego z badanych regionów otrzymano macierz odległości $[d'_{lt}]$ o wymiarach $(t \times t)$, Następnie proponuje się przeprowadzenie analizy miar odległości między wartościami cechy b_1 dla sąsiadujących lat.

3) Periodyzacja poziomu zjawiska ekonomicznego odrębnie dla każdego z badanych regionów

Do wyodrębnienia faz rozwoju zjawiska ekonomicznego wykorzystano metodę taksonomii struktur S. Chomątowskiego i A. Sokołowskiego (1) w jej zmodyfikowanej postaci, umożliwiającej wyodrębnienie grup lat podobnych ze względu na poziom badanej cechy i chronologicznie po sobie następujących.

Do ustalenia tzw. odległości krytycznej wykorzystano formułę przedstawioną poniżej:

$$d^* = \frac{1}{2n} \sum_{r=1}^{n} (\bar{d}_r + \bar{\bar{d}}_r), \tag{13}$$

gdzie: d^* – odległość krytyczna,
$\bar{d}_r, \bar{\bar{d}}_r$ – średnie arytmetyczne z odległości między wartościami cechy statystycznej b_1, r-tego obiektu, pochodzącymi odpowiednio ze wszystkich badanych lat lub z sąsiadujących lat.

Wartości $\bar{d}_r$, $\bar{d}_r$ obliczono w następujący sposób:

$$\bar{d}_r = \frac{2}{t(t-1)} \sum_{t>1}^{t} \sum_{l=1}^{t} d'_{ltr}, \tag{14}$$

$$\bar{d}_r = \frac{1}{t-1} \sum_{l=1}^{t-1} d'_{l,l+1,r}, \tag{15}$$

gdzie: l, $l = 1, \ldots, t$ (numer okresu badania),
$r = 1, \ldots, n$ (numer obiektu badania),
d'_{ltr} – odległość między wartościami cechy b_1 r-tego obiektu, ustalonymi dla l-tego i l-tego okresu badania,
$d'_{l,l+1,r}$ – odległość między wartościami cechy b_1 r-tego obiektu, ustalonymi dla l-tego i $(l+1)$-tego okresu badania.

Kompleksowa analiza porównawcza poziomu zjawiska ekonomicznego wybranych obiektów-regionów może objąć swym zakresem również rejestrację zmian w czasie w ich uporządkowaniu liniowym. Do tego celu można wykorzystać współczynnik korelacji rang C. Spearmana [6], którego przekształconą postać przedstawia następująca formuła:

$$R = 1 - \frac{6 \sum_{r=1}^{n} (z_{rl} - z_{rk})}{n(n^2-1)}, \tag{16}$$

gdzie: z_{rl}, z_{rk} – ranga nadana r-temu obiektowi w uporządkowaniu liniowym, dokonanym ze względu na poziom cechy b_1, w l-tym i l-tym okresie, $r = 1, \ldots, n$ (numer obiektu badania).

Miara ta przyjmuje wartości z przedziału $[-1, 1]$. Wartości bliskie jedności oznaczają bardzo dużą zgodność uporządkowań liniowych obiektów, dokonanych w l-tym i l-tym okresie.

LITERATURA

[1] Chomątowski S., Sokołowski A.: *Taksonomia struktur.* „Przegląd Statystyczny" 1978 nr 2.
[2] Cormack R.M.: *A review of classification (with discussion).* „Journal of the Royal Statistical Society" 1971, vol. 134, part 3, Ser. A.
[3] Fortier J.J., Solomon H.: *Clustering Procedures.* W: *Multivariate Analysis.* Red. P.R. Krishnaiah, New York: Academic Press 1966 nr 62.
[4] Nowak E.: *Wskaźniki podobieństwa wyników podziałów.* „Przegląd Statystyczny" 1985 nr 1.
[5] Walesiak M.: *Metody klasyfikacji w badaniach strukturalnych.* Rozprawa doktorska. Wrocław: AE 1985 (maszynopis).
[6] Zając K.: *Zarys metod statystycznych.* Warszawa: PWE 1982.