PRZEGLĄD STATYSTYCZNY
R. XL – zeszyt 1 – 1993

MAREK WALESIAK

ZAGADNIENIE OCENY PODOBIEŃSTWA ZBIORU OBIEKTÓW W CZASIE W SYNTETYCZNYCH BADANIACH PORÓWNAWCZYCH

Ocenę podobieństwa zbioru obiektów w czasie w syntetycznych badaniach porównawczych przeprowadza się na podstawie wartości cech syntetycznych. W pracy proponuje się dwa mierniki podobieństwa zbioru obiektów w czasie. Stosowanie mierników uzależnione jest od skali pomiaru wartości cech syntetycznych. Konstrukcja pierwszego z nich (wykorzystująca ideę miernika rzędu dokładności prognoz typu ex post H. Theila) zakłada, że wartości cech syntetycznych mierzone są na skali ilorazowej lub przedziałowej. Za pomocą tego miernika mierzy się zmiany w czasie wartości porównywanych cech syntetycznych, a więc mierzy się oddalenie międzyokresowe obiektów. Cenną zaletą miernika jest to, że można go rozłożyć na sumę kilku składników posiadających jasną interpretację, jeśli idzie o rząd i charakter odchyleń wartości porównywanych cech syntetycznych.

Wykorzystując drugi miernik, będący współczynnikiem korelacji tau Kendalla, zakłada się, że wartości cech syntetycznych mierzone są na skali porządkowej. Współczynnik ten pozwala mierzyć stopień podobieństwa dwóch uporządkowań obiektów. Wskazuje więc na stopień przemieszczenia w hierarchii obiektów wraz z upływem czasu.

1

Dany jest niepusty zbiór obiektów A o elementach A_i $(i=1,\ldots,n)$. Niech p_{ir}, p_{is} oznacza wartość cechy syntetycznej, odpowiednio M_r, M_s, dla i-tego obiektu w porównywanych okresach r i s(1).

Wartości cech syntetycznych M_r i M_s są bezpośrednio porównywalne, ponieważ są wyznaczone za pomocą tak samo skonstruowanego syntetycznego miernika rozwoju (SMR), na podstawie tego samego zespołu cech (2). Całe postępowanie,

(1) Narzędziem syntetycznych badań porównawczych jest SMR będący funkcją agregującą znormalizowane wartości cech dla każdego obiektu ze zbioru A. Z formalnego punktu widzenia wartości SMR są realizacjami cechy syntetycznej (por. np. [7], s. 455).

(2) Różne konstrukcje SMR przedstawiono m. in. w pracach [1], [3], [4], [9], [13].

w efekcie którego wyznacza się wartości cech syntetycznych M_r i M_s, jest jednolite dla obu porównywanych okresów. Postępowanie to obejmuje m. in. (por. [13]):

– ujednolicenie charakteru cech będących przedmiotem agregacji poprzez zamiany destymulant i nominant na stymulanty,

– niemianowanie wartości cech i ujednolicenie rzędów ich wielkości poprzez normalizację,

– konstrukcję SMR, w tym ustalenie postaci analitycznej SMR, systemu wag oraz formy wprowadzenia tego systemu do SMR.

W tym miejscu nieco uwagi należy poświęcić tym skalom pomiaru, które będą wykorzystywane w pracy, tj. porządkowej, przedziałowej i ilorazowej(3).

Na wartościach ze skali porządkowej można określić następujące relacje: równości, różności, większości i mniejszości. Dla skali przedziałowej, oprócz relacji właściwych dla skali porządkowej, dopuszcza się relację równości różnic i przedziałów, a dla skali ilorazowej ponadto relację równości stosunków między poszczególnymi wartościami skali.

Dodawanie i odejmowanie dopuszczalne jest dla wartości ze skali przedziałowej. Skala ilorazowa dopuszcza ponadto dzielenie i mnożenie. Jedyną dopuszczalną operacją dla wartości ze skali porządkowej jest zliczanie zdarzeń (tzn. liczby relacji mniejszości, większości i równości jej wartości).

2

Najpierw zostanie przedstawiona konstrukcja miernika podobieństwa zbioru obiektów w czasie, opierającego się na cechach syntetycznych M_r i M_s mierzonych na skali przedziałowej lub ilorazowej.

Zakłada się, że miernik będzie mierzył nie tylko rząd odchyleń od wartości porównywanych cech syntetycznych M_r, M_s, ale również rząd odchyleń będący rezultatem

1° różnicy między średnimi wartościami cech syntetycznych M_r, M_s,

2° różnicy w dyspersji wartości cech syntetycznych M_r, M_s,

3° niezgodności kierunku zmian wartości cech syntetycznych M_r, M_s.

Miernik posiadający wszystkie te cechy przyjmuje następującą postać

$$P^2(M_r, M_s) = P_{rs}^2 = \frac{1}{n} \sum_{i=1}^{n} (p_{ir} - p_{is})^2 . \tag{1}$$

Miernik (1) przyjmuje wartość 0 w przypadku, gdy nie ma żadnych różnic w wartościach cech syntetycznych M_r i M_s. Pierwiastek kwadratowy z wyrażenia (1) informuje, jaki jest przeciętny rząd odchyleń wartości porównywanych cech syntetycznych z okresów r i s.

---

(3) Definicje skal pomiaru oraz ich szersze omówienie w aspekcie syntetycznych badań porównawczych przedstawiono w pracy [13].

Wyrażoną wzorem (1) wielkość można rozłożyć na sumę trzech składników

$$P_{rs}^2 = P_1^2 + P_2^2 + P_3^2, \tag{2}$$

pozwalających określić bliżej „rząd" i „charakter" różnic w wartościach cech syntetycznych M_r i M_s.

Mierniki cząstkowe P_1^2, P_2^2 i P_3^2 (niosące informacje, o których mowa w punktach 1°, 2° i 3°) określają wzory([4])

$$P_1^2 = (\bar{p}_{.r} - \bar{p}_{.s})^2, \tag{3}$$

$$P_2^2 = (S_r - S_s)^2, \tag{4}$$

$$P_3^2 = 2 S_r S_s (1 - \rho), \tag{5}$$

gdzie $\bar{p}_{.r}$, $S_r(\bar{p}_{.s}, S_s)$ to, odpowiednio, średnia arytmetyczna i odchylenie standardowe wartości r-tej (s-tej) cechy syntetycznej, ρ – współczynnik korelacji liniowej Pearsona między wektorami $\mathbf{p}_{.s} = (p_{1s}, \ldots, p_{ns})$ i $\mathbf{p}_{.r} = (p_{1r}, \ldots, p_{nr})$.

3

Jeśli świadomie zdecydujemy się na utratę informacji i potraktujemy otrzymane wartości p_{ir} i p_{is} cech syntetycznych M_r i M_s, tak jakby były one mierzone na skali porządkowej, to stosując współczynnik tau Kendalla możemy ocenić podobieństwo uporządkowań zbioru obiektów z okresów r i s([5]). Współczynnik ten wyrażający skorelowanie cech mierzonych na skali porządkowej pozwala określić stopień zmiany uporządkowania obiektów wraz z upływem czasu.

Zaproponowany przez M. G. Kendalla współczynnik tau wyraża się wzorem ([5], s. 19; [12])

$$K_{rs} = \frac{\sum_{j=2}^{n} \sum_{i=1}^{j-1} a_{ij} b_{ij}}{\left(\sum_{j=2}^{n} \sum_{i=1}^{j-1} a_{ij}^2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} b_{ij}^2 \right)^{0,5}}, \tag{6}$$

gdzie

$$a_{ij}(b_{ij}) = \begin{cases} 1, & \text{jeśli } p_{ir} > p_{jr} \quad (p_{is} > p_{js}), \\ 0, & \text{jeśli } p_{ir} = p_{jr} \quad (p_{is} = p_{js}), \\ -1, & \text{jeśli } p_{ir} < p_{jr} \quad (p_{is} < p_{js}). \end{cases} \tag{7}$$

Współczynnik korelacji K_{rs} przyjmuje wartości z przedziału $[-1; 1]$. Wartość 1 oznacza pełną zgodność uporządkowań, wartość -1 natomiast pełną ich przeciwstawność.

([4]) Rozbicie wzoru (1) na trzy składniki zaczerpnięte zostało z wzoru H. Theila na miernik rzędu dokładności prognozy typu ex post (por. [10], s. 119; [14], s. 184).

([5]) Strata informacji polega na przejściu z wyższego poziomu (skala ilorazowa lub przedziałowa) na niższy poziom pomiaru (skala porządkowa).

Można zadać pytanie, dlaczego w pracy preferuje się współczynnik tau Kendalla, a nie powszechnie znany i stosowany współczynnik korelacji rang Spearmana. Współczynnik korelacji rang Spearmana to w pewien sposób transformowany współczynnik korelacji liniowej Pearsona. W tej transformacji wykorzystuje się specyfikę kolejnych n liczb naturalnych (por. np. [11], s. 160-162). Współczynnik ten nie jest typową miarą korelacji rang, bowiem stosując go zakłada się, że odległości pomiędzy sąsiednimi wartościami na skali porządkowej są sobie równe (na skali porządkowej odległości między dowolnymi dwiema wartościami nie są znane).

W pracy [12] zwrócono uwagę na nie dostrzegany w polskiej literaturze statystycznej fakt, że współczynnik korelacji tau Kendalla (a nie współczynnik rang Spearmana) jest dla wyników pomiaru porządkowego szczególną postacią współczynnika korelacji liniowej Pearsona([6]).

4

Miernik o postaci (1) zostanie zastosowany do oceny zmian w wartościach cechy syntetycznej wyrażającej warunki mieszkaniowe ludności miejskiej województwa jeleniogórskiego w okresie 1978-1988. Zmiany w hierarchii miast ocenione zostaną za pomocą współczynnika (6).

Zbiór obiektów obejmuje 25 miast województwa jeleniogórskiego (w podziale administracyjnym z 6 grudnia 1988 r.) oraz, dla stworzenia szerszej bazy porównawczej, trzy dalsze obiekty, tj. województwo, miasta, gminy (są to obiekty o umownych nazwach, które prezentują przeciętne wartości cech dla ogółu jednostek populacji).

Dla scharakteryzowania obiektów przyjęto zestaw 11 cech określających poziom warunków mieszkaniowych ludności (dane statystyczne opracowane na podstawie [2] zawiera praca [8]):

$x^D_{\cdot 1t}$ – przeciętna liczba osób w mieszkaniu,
$x^D_{\cdot 2t}$ – przeciętna liczba osób na izbę,
$x_{\cdot 3t}$ – przeciętna powierzchnia użytkowa mieszkania w m^2,
$x_{\cdot 4t}$ – przeciętna powierzchnia użytkowa mieszkania na 1 osobę w m^2,
$x^D_{\cdot 5t}$ – przeciętna liczba gospodarstw domowych przypadających na 1 mieszkanie,
$x_{\cdot 6t}$ – odsetek mieszkań wyposażonych w wodociąg,
$x_{\cdot 7t}$ – odsetek mieszkań wyposażonych w ustęp spłukiwany,
$x_{\cdot 8t}$ – odsetek mieszkań wyposażonych w łazienkę,
$x_{\cdot 9t}$ – odsetek mieszkań wyposażonych w ciepłą bieżącą wodę,
$x_{\cdot 10t}$ – odsetek mieszkań wyposażonych w gaz sieciowy,
$x_{\cdot 11t}$ – odsetek mieszkań wyposażonych w CO.

Ze względu na to, że wśród cech występują destymulanty (cechy $x^D_{\cdot 1t}$, $x^D_{\cdot 2t}$, $x^D_{\cdot 5t}$) i stymulanty (pozostałe cechy) oraz na to, że cechy posiadają różne miana

([6]) Odpowiedni dowód zawarty jest m. in. w pracy [12].

i rzędy wielkości, niemożliwa jest agregacja bez przeprowadzenia postępowania unifikacyjnego. Ujednolicenia charakteru cech dokonamy przekształcając destymulanty na stymulanty według wzoru (por. [13]):

$$x_{ikt}=c(x_{ikt}^{D})^{-1}, \quad c>0 \tag{8}$$

gdzie x_{ikt} – wartość k-tej cechy w i-tym obiekcie w roku t ($i=1,\ldots,28$; $k=1,\ldots,11$, $t=r, s$; „r" – rok 1988, „s" – rok 1978), c – stała przyjmowana arbitralnie (w pracy przyjęto $c=1$).

Ujednolicenie mian oraz rzędów wielkości cech uzyskuje się przez ich normalizację. Wszystkie cechy mierzone są na skali ilorazowej, zatem jako formułę normalizacji można wykorzystać (por. [13]):

$$z_{ikt}=x_{ok.}^{-1}x_{ikt}, \tag{9}$$

gdzie z_{ikt} – k-ta wartość w i-tym obiekcie w roku t przekształconej (znormalizowanej) cechy, a liczba $x_{ok.}$ oznacza podstawę normalizacji k-tej cechy, którą można różnie definiować (por. np. [13]).

Dla zachowania porównywalności w czasie za podstawę przyjęto korzystniejsze wartości cech z lat 1978, 1988 odpowiadające obiektowi o nazwie „województwo" (wszystkie tak określone wartości cech pochodziły z roku 1988)(7). Uporządkowanie miast ze względu na poziom warunków mieszkaniowych ludności w latach 1978 i 1988 (por. tab. 1) ustalono za pomocą SMR o postaci

$$p_{it}=\frac{\sum_{k=1}^{m}\alpha_k z_{ikt}}{\sum_{k=1}^{m}\alpha_k}, \tag{10}$$

gdzie p_{it} – wartość SMR dla i-tego obiektu w okresie t (i-ta wartość cechy syntetycznej w okresie t).

W pracy przyjęto, że wszystkie cechy są jednakowo ważne, zatem $\alpha_k=1$ ($k=1,\ldots,m=11$).

Obliczony na podstawie tab. 1 miernik (1) i mierniki cząstkowe (3), (4) i (5) wynoszą

$$P_{rs}^2=0{,}028929,$$

$$P_1^2=0{,}027497, \quad P_2^2=0{,}000563, \quad P_3^2=0{,}000869,$$

przy czym

$$\bar{p}_{.r}=0{,}996, \quad \bar{p}_{.s}=0{,}830, \quad S_r=0{,}112, \quad S_s=0{,}136, \quad \rho=0{,}972.$$

---

(7) W analizowanym przykładzie $x_{ok.}=\max_t\{x_{okt}\}$, gdzie o jest numerem ustalonego obiektu (województwo). Dla zachowania porównywalności w czasie wartości cech syntetycznych M_r i M_s w badaniach dynamicznych podstawa normalizacji k-tej cechy musi przyjmować taką samą wartość liczbową dla każdego okresu.

Tabela 1

Wartości cechy syntetycznej wyrażającej poziom warunków mieszkaniowych ludności miejskiej województwa jeleniogórskiego w latach 1978 i 1988

Lp.	Wyszczególnienie	p_{ir}	p_{is}
1	Zgorzelec	1,202	1,123
2	Lwówek Śląski	1,187	1,058
3	Bolesławiec	1,174	1,074
4	Jelenia Góra	1,144	0,999
5	Gryfów Śląski	1,103	0,929
6	Szklarska Poręba	1,090	0,926
7	Miasta	1,084	0,940
8	Karpacz	1,081	0,922
9	Lubań	1,072	0,919
10	Bolków	1,066	0,907
11	Kowary	1,043	0,887
12	Piechowice	1,039	0,855
13	Kamienna Góra	1,022	0,869
14	Województwo	1	0,831
15	Lubawka	0,981	0,806
16	Wleń	0,973	0,833
17	Nowogrodziec	0,962	0,710
18	Świeradów Zdrój	0,936	0,719
19	Lubomierz	0,935	0,742
20	Świerzawa	0,912	0,700
21	Bogatynia	0,908	0,777
22	Leśna	0,882	0,670
23	Mirsk	0,875	0,666
24	Węgliniec	0,867	0,669
25	Zawidów	0,860	0,723
26	Wojcieszów	0,838	0,649
27	Pieńsk	0,835	0,722
28	Gminy	0,816	0,619

Źródło: [8].

Przeciętne odchylenie wartości cech syntetycznych z lat 1978 i 1988 wyniosło 0,170. Było to wynikiem głównie wzrostu średniego poziomu wartości cechy syntetycznej. Zanotowano wysoką zgodność kierunku zmian wartości cech syntetycznych z porównywanych okresów oraz niewielki spadek w zróżnicowaniu wartości cechy syntetycznej świadczący o zmniejszeniu (choć nieznacznym) dysproporcji między miastami pod względem poziomu rozwoju warunków mieszkaniowych ludności.

Na podstawie współczynnika $K_{rs}=0{,}857$ można wnioskować o niewielkich zmianach w uporządkowaniu miast pod względem poziomu warunków mieszkaniowych ludności.

Akademia Ekonomiczna we Wrocławiu, Wydział Gospodarki Regionalnej i Turystyki w Jeleniej Górze

LITERATURA

[1] Borys T., *Kategoria jakości w statystycznej analizie porównawczej*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 284, Seria: Monografie i opracowania nr 23, Wrocław 1984.

[2] *Charakterystyka zmian demograficzno-społecznych ludności i warunków mieszkaniowych w latach 1978-1988 – województwo jeleniogórskie*, Narodowy Spis Powszechny z dnia 6 grudnia 1988 r., GUS, Warszawa 1989.

[3] Grabiński T., *Wielowymiarowa analiza porównawcza w badaniach dynamiki zjawisk ekonomicznych*, Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, Seria specjalna: Monografie nr 61, Kraków 1984.

[4] Hellwig Z., *Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju i strukturę wykwalifikowanych kadr*, Przegląd Statystyczny 4 (1968), s. 307-327.

[5] Kendall M. G., *Rank Correlation Methods*, Griffin, London 1955.

[6] Kendall M. G., Buckland W. R., *Słownik terminów statystycznych*, PWE, Warszawa 1986.

[7] Nowak E., *Badanie zgodności metod konstruowania taksonomicznych mierników rozwoju*, Przegląd Statystyczny 3-4 (1982), s. 455-463.

[8] Obrębalski M., Walesiak M., *Pomiar i identyfikacja zmian poziomu warunków mieszkaniowych ludności miejskiej regionu jeleniogórskiego w latach 1978-1988*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 600, Wrocław 1991, s. 39-50.

[9] Ostasiewicz W., *Zastosowanie miary rozmytej do porównań syntetycznych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 190, Wrocław 1981, s. 79-102.

[10] Pawłowski Z., *Prognozy ekonometryczne*, PWN, Warszawa 1973.

[11] Steczkowski J., Zeliaś A., *Statystyczne metody analizy cech jakościowych*, PWE, Warszawa 1981.

[12] Walesiak M., *O stosowalności miar korelacji w analizie wyników pomiaru porządkowego*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 600, Wrocław 1991, s. 13-19.

[13] Walesiak M., *Syntetyczne badania porównawcze w świetle teorii pomiaru*, Przegląd Statystyczny 1-2 (1990), s. 37-46.

[14] Zeliaś A., *Teoria prognozy*, PWE, Warszawa 1984.

Praca wpłynęła do Redakcji w sierpniu 1990 r.
Wersja ostateczna w październiku 1991 r.

ПРОБЛЕМА ОЦЕНКИ УРОВИЯ ПОДОБИЯ МНОЖЕСТВА ОБЪЕКТОВ ВО ВРЕМЕНИ

Резюме

Анализ уровия подобия множества объектов во времени проводится с использованием синтетических показателей. Предложены две меры подобия, которых применение существенно зависит от шкалы измерений. Построение первой из них (использующее идею точности прогнозов экс пост Х. Тайпа) предполагает использование чистой или интервальной шкалы измерения. С ее помощью можно исследовать изменение во времени синтетических показателей, т.е. установить степень промежуточного расстаяния объектоб.

Вторая мера является корреляционным коэффициентом тау Кендалла. Она предполагает существование порядковой шкалы измерения и позваляет вычислять уровень подобия двух упорядоченных объектоб.

ON THE PROBLEM OF SIMILARITY OF SETS OF THE SAME OBJECTS IN TIME IN SYNTHETIC COMPARATIVE STUDIES

Summary

Similarity of sets of the same objects in time is analyzed on the basis of values of synthetic variables. In the paper there are proposed two measures of similarity. Application of the measures depends on scales of measurement of synthetic variables.

The first measure (employing the idea of Theil's measure of ex post precission of forecasts) is based on synthetic variables being measured on relative or interval scales. Using it the difference between the variables, i.e. the distance between objects is quantify.

The second measure (being the Kendall's tau correlation coefficient) may be used for synthetic variables measured on order scale. The coefficient is used to quantify the similarity between orderings of sets of objects.