B A D A N I A O P E R A C Y J N E I D E C Y Z J E
Nr 1 1993

Marek WALESIAK*

STRATEGIE POSTĘPOWANIA W BADANIACH STATYSTYCZNYCH W PRZYPADKU ZBIORU ZMIENNYCH MIERZONYCH NA SKALACH RÓŻNEGO TYPU**

Omówiono strategie postępowania w badaniach statystycznych w przypadku zbioru zmiennych mierzonych na skalach różnego typu w odniesieniu do zagadnień klasyfikacji i porządkowania liniowego. Zwrócono uwagę na strategię, która wymaga zastosowania narzędzi statystycznych właściwych dla skali porządkowej. Zaproponowano konstrukcję miary odległości oraz syntetycznego miernika rozwoju dla obiektów opisanych zmiennymi porządkowymi.

1.

Głównym celem klasyfikacji jest poznanie natury obiektów (por. praca [3]), tzn. badanie podobieństwa lub odrębności obiektów i ich zbiorów. Celem tym jest więc podział zbioru obiektów na klasy, zawierające obiekty podobne pod względem wartości zmiennych, wyrażających naturę obiektów.

Zadaniem metod porządkowania liniowego zbioru obiektów jest uszeregowanie, czyli ustalenie kolejności obiektów lub ich zbiorów według określonego kryterium. Metody te mogą być zatem stosowane wtedy, gdy można przyjąć pewne nadrzędne kryterium, według którego można uporządkować obiekty od „najlepszego" do „najgorszego".

Narzędziem metod klasyfikacji są różnego rodzaju miary podobieństwa obiektów, metod porządkowania liniowego zaś syntetyczny miernik rozwoju (SMR), będący pewną funkcją agregującą informacje cząstkowe zawarte w poszczególnych zmiennych i wyznaczoną dla każdego obiektu ze zbioru obiektów A. Stosowanie

* Wydział Gospodarki Regionalnej i Turystyki Akademii Ekonomicznej we Wrocławiu, ul. Nowowiejska 3, 58–500 Jelenia Góra.

** Praca została wykonana w ramach grantu KBN 09139101 nt. *Statystyczna klasyfikacja wielowymiarowa. Teoria i jej zastosowanie ekonomiczne.*

konkretnych miar podobieństwa w klasyfikacji i różnych konstrukcjach SMR jest uzależnione od skal pomiaru zmiennych.

W artykule zwrócono szczególną uwagę na strategię postępowania w badaniach statystycznych w przypadku zbioru zmiennych, który wymaga zastosowania narzędzi statystycznych właściwych dla skali porządkowej. Nie ma bowiem w literaturze statystycznej propozycji miar podobieństwa ani konstrukcji syntetycznych mierników rozwoju dla obiektów opisanych zmiennymi mierzonymi na tej skali, o czym wspomniano m.in. w opracowaniach [2], [5], [7], [11].

Problematyka poruszona w artykule wymaga wprowadzenia podstawowych pojęć z dziedziny teorii pomiaru.

Przez pomiar rozumie się przyporządkowanie liczb obiektom zgodnie z określonymi regułami w taki sposób, aby liczby odzwierciedlały zachodzące między tymi obiektami relacje (por. np. prace [10], s. 54; [4], s. 17).

Podstawą teorii pomiaru jest pojęcie skali.

Definicja 1 (por. [1], s. 101–102; [15], s. 37). Taką uporządkowaną czwórkę $U = \langle A; G; H; F \rangle$, gdzie:

a) A – to niepusty zbiór obiektów, H – zbiór liczb rzeczywistych, G – klasa funkcji odwzorowujących A w H, F – klasa funkcji odwzorowujących H w H,

b) dla wszystkich $g \in G$ oraz $f \in F$, $f \cdot g \in G$,

c) F zawiera przekształcenie H na H, ponadto dla każdego f_k, $f_l \in F$ złożenie $f_k \cdot f_l \in F$,

nazywa się skalą pomiaru.

W teorii pomiaru rozróżnia się 4 podstawowe skale pomiaru, wprowadzone przez Stevensa [13].

Definicja 2 (por. [1], s. 103; [16], s. 13–14). $U = \langle A; G; H; F \rangle$ jest skalą nominalną wtedy i tylko wtedy, gdy F jest zbiorem wszystkich funkcji f odwzorowujących H w H $(H = R)$ takich, że

$$f - \text{funkcja wzajemnie jednoznaczna} \tag{1}$$

Definicja 3 (por. [1], s. 103; [16], s. 14). $U = \langle A; G; H; F \rangle$ jest skalą porządkową wtedy i tylko wtedy, gdy F jest zbiorem wszystkich funkcji f odwzorowujących H w $H (H = R)$ takich, że

$$f - \text{funkcja ściśle monotonicznie rosnąca} \tag{2}$$

Definicja 4 (por. [1], s. 103; [15], s. 37). $U = \langle A; G; H; F \rangle$ jest skalą interwałową (przedziałową) wtedy i tylko wtedy, gdy H jest zbiorem wszystkich liczb rzeczywistych R i F jest zbiorem funkcji f takich, że dla dodatniego b

$$f(y) = by + a, \quad f(y) \in R \tag{3}$$

dla wszystkich $y \in R$.

Definicja 5 (por. [1], s. 103; [15], s. 38). $U = \langle A; G; H; F \rangle$ jest skalą ilorazową (stosunkową) wtedy i tylko wtedy, gdy H jest zbiorem liczb rzeczywistych dodatnich

R_+ i F jest zbiorem funkcji f takich, że dla dodatniego b

$$f(y) = by, \quad f(y) \in R_+ \tag{4}$$

dla wszystkich $y \in R_+$.

Skale te są uporządkowane od najsłabszej (nominalna) aż do najmocniejszej (ilorazowa). Wynika to z definicji 6.

Definicja 6 (por. [14], s. 52). Skala U_2 jest mocniejsza od skali U_1 zawsze i tylko wtedy, gdy jej dopuszczalne przekształcenie jest zdegenerowanym przypadkiem dopuszczalnego przekształcenia skali U_1.

Na wartościach poszczególnych skal, ze względu na dopuszczalne przekształcenie, można wyznaczać następujące relacje:

a) skala nominalna – relacje: równości, różności,

b) skala porządkowa – relacje: równości, różności, większości, mniejszości,

c) skala przedziałowa – relacje: równości, różności, mniejszości, większości, równości różnic i przedziałów,

d) skala ilorazowa – relacje: równości, różności, mniejszości, większości, równości różnic i przedziałów, równości stosunków między poszczególnymi wartościami skali.

Wykonywanie operacji arytmetycznych dodawania i odejmowania jest dopuszczalne na wartościach skali przedziałowej. Skala ilorazowa dopuszcza ponadto wykonywanie na wartościach skali operacji dzielenia i mnożenia. Jedyną dopuszczalną operacją empiryczną na wartościach skali nominalnej i porządkowej jest zliczanie zdarzeń (tzn. ile relacji mniejszości, większości i równości określono na wartościach np. skali porządkowej).

Jedna z podstawowych reguł teorii pomiaru mówi, że jedynie wyniki pomiaru w skali mocniejszej mogą być transformowane na liczby należące do skali słabszej (por. np. prace [12], s. 17; [15], s. 40). Stosując zaś dozwolone przekształcenie wartości na skali, zachowujemy niezmienność typu skali przyjętej dla danej zmiennej.

Typ skali, ze względu na dopuszczalne przekształcenia, determinuje stosowalność rozmaitych technik statystyczno-ekonometrycznych.

Definicja 7 (por. np. [14], s. 61). Technikami statystycznymi dopuszczalnymi dla danego typu skali są takie techniki, które dostarczają wyników (w sensie relacji) niezmiennych względem dopuszczalnych przekształceń.

2.

W zagadnieniu klasyfikacji w zbiorze mogą być zmienne mierzone na różnych skalach pomiaru (czyli może wystąpić tzw. mieszanka zmiennych), z kolei porządkowanie liniowe wymaga, aby w zbiorze były zmienne mierzone przynajmniej na skali porządkowej (z uwagi na to, że porządkowanie obiektów staje się możliwe, gdy dopuszczalne jest określenie na wartościach zmiennych relacji większości i mniejszości).

Problem stosowania konkretnych konstrukcji miar podobieństwa w klasyfikacji nie występuje wtedy, gdy wszystkie zmienne są mierzone na skali: a) przedziałowej i (lub) ilorazowej, b) nominalnej. Wynika to z faktu, że dla tych skal istnieją rozmaite konstrukcje miar podobieństwa. Bardzo dobry przegląd miar podobieństwa dla tych grup zmiennych przedstawił Anderberg [2]. Jeśli idzie o porządkowanie liniowe, to opracowano wiele konstrukcji SMR w przypadku, gdy w zbiorze znajdują się zmienne mierzone tylko na skali przedziałowej i (lub) ilorazowej. Różne konstrukcje SMR dla tych grup zmiennych przedstawił m.in. Walesiak w pracy [15].

Nie wypracowano dotychczas w literaturze statystycznej miar podobieństwa obiektów ani konstrukcji SMR, które można by stosować w sytuacji, gdy w zbiorze są zmienne mierzone tylko na skali porządkowej. Celem prezentowanego artykułu jest m.in. uzupełnienie tej luki.

W konstrukcji miary odległości obiektów opisanych zmiennymi porządkowymi wykorzystano ideę współczynnika korelacji zmiennych porządkowych (tau) Kendalla (por. [8], s. 19; [16]).

Dany jest niepusty zbiór obiektów A opisanych m zmiennymi porządkowymi. Z uwagi na to, że na skali porządkowej dopuszczalną operacją empiryczną jest tylko zliczanie zdarzeń (tzn. wyznaczanie liczby relacji większości, mniejszości i równości), proponuje się następującą konstrukcję miary odległości:

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj}\, b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj}\, b_{klj}}{2\left[\left(\sum_{j=1}^{m} a_{ikj}^2 + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj}^2\right)\left(\sum_{j=1}^{m} b_{kij}^2 + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} b_{klj}^2\right)\right]^{1/2}} \tag{5}$$

gdzie:

$$a_{ipj}\,(b_{krj}) = \begin{cases} 1, \text{ jeżeli } x_{ij} > x_{pj}\ (x_{kj} > x_{rj}) \\ 0, \text{ jeżeli } x_{ij} = x_{pj}\ (x_{kj} = x_{rj}) \\ -1, \text{ jeżeli } x_{ij} < x_{pj}\ (x_{kj} < x_{rj}) \end{cases} \tag{6}$$

$p = k, l;\ r = i, l,$

$i, k, l = 1, \ldots, n$ – numer obiektu,

$j = 1, \ldots, m$ – numer zmiennej porządkowej,

$x_{ij}(x_{kj}, x_{lj})$ – i-ta (k-ta, l-ta) obserwacja na j-tej zmiennej porządkowej,

$\sum_{j=1}^{m} a_{ikj}^2 + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj}^2$ – liczba relacji większości i mniejszości określona dla obiektu i,

$\sum_{j=1}^{m} b_{kij}^2 + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} b_{klj}^2$ – liczba relacji większości i mniejszości określona dla obiektu k.

Miara odległości d_{ik} przyjmuje wartości z przedziału [0; 1]. Wartość 0 oznacza, że dla porównywanych obiektów i, k między odpowiadającymi sobie obserwacjami na zmiennych porządkowych zachodzą tylko relacje równości. Z kolei wartość 1 przyjmuje wtedy, gdy dla porównywanych obiektów i, k między odpowiadającymi sobie obserwacjami na zmiennych porządkowych zachodzą tylko relacje większości (mniejszości) lub relacje większości (mniejszości) oraz relacje równości i relacje te są zachowane w stosunku do pozostałych obiektów (a więc obiektów o numerach $l = 1, \ldots, n$; gdzie $l \neq i, k$).

Jeśli we wzorach (5) i (6) w miejsce indeksu k wstawimy indeks w (oznaczający numer obiektu – wzorca), to otrzymamy konstrukcję SMR oznaczaną (5') i (6'). W takiej sytuacji miara (5') oznacza odległość obiektu i-tego od obiektu – wzorca w.

Sytuacja komplikuje się, jeśli w zbiorze znajdują się zmienne mierzone na skalach różnych rodzajów. Na podstawie literatury przedmiotu (por. [7], s. 32–37; [9], [5], s. 25–27) do rozwiązania tego problemu można wykorzystać następujące sposoby:

a) Przeprowadzić klasyfikację i porządkowanie liniowe zbioru obiektów osobno dla każdej grupy zmiennych. Jeśli otrzymane w ten sposób wyniki są w miarę zgodne, to problem można uznać za rozwiązany. Sytuacja komplikuje się wtedy, gdy wyniki znacznie od siebie odbiegają.

b) Wykorzystać w analizie tylko zmienne jednego ustalonego typu (dominującego w zbiorze zmiennych) odrzucając zmienne innego typu. Wyniki otrzymane na podstawie zbioru zmiennych, uzyskanego w taki sposób, są na ogół bardzo zniekształcone (z uwagi na to, że musimy zrezygnować z części informacji, jakie niosą odrzucone zmienne).

c) Praktycznie zaniedbać fakt, że zmienne są mierzone na skalach różnych typów i stosować metody właściwe dla zmiennych jednego typu. Zazwyczaj traktuje się zmienne nominalne i porządkowe tak, jak przedziałowe i ilorazowe i stosuje się techniki właściwe tym skalom. Sposób ten, choć atrakcyjny z aplikacyjnego punktu widzenia, jest nie do przyjęcia ze względów metodologicznych (następuje tutaj bowiem sztuczne wzmocnienie skali pomiaru).

d) Dokonać transformacji zmiennych tak, by sprowadzić je do skali jednego typu. Podstawowa reguła pomiaru mówi, że jedynie wyniki pomiaru w skali mocniejszej mogą być transformowane na liczby należące do skali słabszej. Wynika stąd, że należy przekodować wszystkie obserwacje na zmiennych na pomiary na skali najsłabszej. Operacji tej towarzyszy jednak utrata informacji. Proponowane są również procedury wzmacniania skal pomiaru (por. prace [2], [11]). Są to aproksymacyjne metody przekształcania skal słabszych w silniejsze, opierające się na pewnych dodatkowych informacjach. Z punktu widzenia teorii pomiaru wzmac-

nianie skal jest jednak niemożliwe, ponieważ z mniejszej ilości informacji nie można uzyskać większej.

e) Posłużyć się metodami (miarami podobieństwa, konstrukcjami SMR) dopuszczającymi stosowanie zmiennych mierzonych na różnych skalach. Sposobu tego nie da się praktycznie wykorzystać ze względu na brak takich miar podobieństwa i konstrukcji SMR. Wprawdzie Gower [6], a następnie Kaufman i Rousseeuw [7] zaproponowali taką miarę podobieństwa obiektów, ale w świetle teorii pomiaru wątpliwe są ich podstawy konstrukcyjne.

Dotychczas w empirycznych zastosowaniach klasyfikacji i porządkowania liniowego, gdy w zbiorze zmiennych występowały zmienne mierzone co najmniej na skali porządkowej, korzystano ze sposobu c), w którym zmienne porządkowe traktowano jak zmienne przedziałowe lub ilorazowe. Zaproponowane w artykule miary: odległości obiektów o postaci (5) i SMR o postaci (5') pozwalają na stosowanie – zgodnego z teorią pomiaru – sposobu d), w którym obserwacje na zmiennych przedziałowych i ilorazowych zostają przekodowane na pomiary na zmiennych porządkowych.

Szczególna przydatność miar (5) i (5') przejawia się w badaniach marketingowych, w których często w zbiorze zmiennych występują zmienne porządkowe.

Bibliografia

[1] ADAMS E.W., FAGOT R.F., ROBINSON R.E., *A theory of appropriate statistics*, Psychometrika 1965 (30), 90–127.

[2] ANDERBERG M.R., *Cluster analysis for applications*, Academic Press, New York, San Francisco, London 1973.

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

[4] CHOYNOWSKI M., *Pomiar w psychologii* [w:] Problemy psychologii matematycznej, red. J. Kozielecki, PWN, Warszawa 1971.

[5] GORDON A.D., *Classification*, Chapman and Hall, London 1981.

[6] GOWER J.C., *A general coefficient of similarity and some of its properties*, Biometrics 1971 (27), 857–874.

[7] KAUFMAN L., ROUSSEEUW P.J., *Finding groups in data: an introduction to cluster analysis*, Wiley, New York 1990.

[8] KENDALL M.G., *Rank correlation methods*, Griffin, London 1955.

[9] KOLONKO J., *O wykorzystaniu w badaniach taksonomicznych danych pierwotnych mierzonych na skalach różnego typu*, Materiały konferencyjne nt. „Metody taksonomiczne i ich zastosowanie w badaniach ekonomicznych", Szklarska Poręba 25.10.1979 r.

[10] PAWŁOWSKI T., *Metodologiczne zagadnienia humanistyki*, PWN, Warszawa 1969.

[11] POCIECHA J., *Statystyczne metody segmentacji rynku*, Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, Seria specjalna: Monografie nr 71, Kraków 1986.

[12] STECZKOWSKI J., ZELIAŚ A., *Statystyczne metody analizy cech jakościowych*, PWE, Warszawa 1981.

[13] STEVENS S.S., *Measurement, psychophysics and utility* [w:] C.W. Churchman, P. Ratoosh (eds.), Measurement; definitions and theories, Wiley, New York 1959.

[14] WALENTA K., *Podstawowe pojęcia teorii pomiaru* [w:] J. Kozielecki, Problemy psychologii matematycznej, PWN, Warszawa 1971.
[15] WALESIAK M., *Syntetyczne badania porównawcze w świetle teorii pomiaru*, Przegląd Statystyczny z. 1–2, 1990, 37–46.
[16] WALESIAK M., *O stosowalności miar korelacji w analizie wyników pomiaru porządkowego*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, nr 600, 13–19, Wrocław 1991.
[17] WALESIAK M., *Zagadnienie oceny podobieństwa zbioru obiektów w czasie w syntetycznych badaniach porównawczych*, Przegląd Statystyczny z. 1, 1993.

Strategies used in statistical studies in the case of variables measured in different scales

Strategies to be used in statistical studies, particularly for classification and ordering methods when the variables are measured in different scales are discussed.

Attention is paid to the case of variables measured in ordinal scale. As was pointed out by Anderberg [2], Gordon [5], Pociecha [11], Kaufman and Rousseeuw [7], for these variables there are no proposals as far as similarity measures and synthetic measures are concerned.

Some proposals in this area are given in the present paper. They are based on Kendall's rank correlation coefficient.

Verified by Marzena Łuczkiewicz