PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 — 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Artur Zaborski

Uniwersytet Ekonomiczny we Wrocławiu

MOŻLIWOŚCI UNIKNIĘCIA ZDEGENEROWANYCH ROZWIĄZAŃ W ANALIZIE *UNFOLDING* PRZY WYKORZYSTANIU ALGORYTMU PREFSCAL

1. Wstęp

Analiza *unfolding* jest metodą skalowania wielowymiarowego najczęściej wykorzystywaną w badaniach preferencji. Jej celem jest odkrycie, na podstawie danych preferencji, wspólnej przestrzeni punktów reprezentujących respondentów i badane obiekty, która pozwala na ocenę zależności występujących pomiędzy respondentami, obiektami oraz pomiędzy respondentami a obiektami (w zależności od celu badania możliwa jest również demonstracja punktów reprezentujących respondentów i zmienne lub obiekty i zmienne).

Jeżeli dane preferencji mierzone są na skali porządkowej, istnieje ryzyko, że w wyniku zastosowania procedur skalowania wielowymiarowego otrzymamy zdegenerowane, nieinterpretowalne rozwiązanie. W takim przypadku prawdziwa struktura danych pozostaje nadal ukryta, mimo że bliska zeru wartość funkcji dopasowania wskazuje pozornie na doskonałe dopasowanie konfiguracji punktów do danych wejściowych.

W literaturze przedmiotu proponuje się różne metody, których celem jest uniknięcie trywialnych rozwiązań (zob. [Borg, Groenen 2005, s. 317-330]). Wszystkie można podzielić na trzy grupy, polegające na modyfikacji:
- danych wejściowych,
- transformacji monotonicznej,
- funkcji dopasowania.

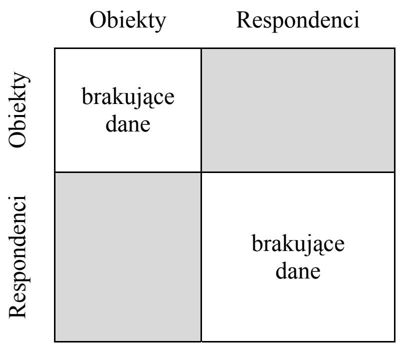
Celem artykułu jest przedstawienie modelu PREFSCAL, w którym uniknięcie rozwiązań trywialnych jest możliwe dzięki modyfikacji funkcji dopasowania STRESS, przez wykorzystanie w jej konstrukcji współczynnika zmienności Pearsona.

W artykule dokonano ponadto porównania wyników badania uzyskanego za pomocą analizy *unfolding* dla różnych wartości parametrów nakładanych na funkcję dopasowania w celu znalezienia rozwiązania optymalnego. Analizę przeprowadzono z wykorzystaniem pakietu statystycznego *SPSS*.

2. Podstawy analizy *unfolding*

Analiza *unfolding* jest metodą, w której na podstawie indywidualnych ocen preferencji dokonuje się rozmieszczenia na mapie percepcyjnej punktów reprezentujących respondentów oraz badane obiekty. Pozycja na mapie percepcyjnej, jaką zajmują punkty oznaczające respondentów lub grupy respondentów, to tzw. punkty idealne, czyli punkty reprezentujące obiekt posiadający najbardziej preferowaną przez respondenta kombinację zmiennych. Pozycja punktów reprezentujących obiekty w stosunku do punktu idealnego określa względne preferencje respondentów, tzn. ich odległości od punktu idealnego mają takie samo uporządkowanie jak rangowe uporządkowanie preferencji.

W analizie *unfolding* dane wejściowe przedstawione są w postaci prostokątnej macierzy preferencji. Wiersze tej macierzy różnią się od jej kolumn. Pierwsze reprezentują respondentów, drugie – obiekty. Elementy poszczególnych wierszy macierzy są ocenami preferencji dla każdego z respondentów (najczęściej otrzymuje się je w wyniku rangowania). Macierz preferencji można traktować jako podmacierz macierzy niepodobieństw, w której dane są tylko niepodobieństwa między respondentami a obiektami (rys. 1).



Rys. 1. Macierz preferencji jako podmacierz macierzy odległości
Źródło: opracowano na podstawie pracy [Borg, Groenen 2005, s. 295].

Taką macierz wykorzystuje się do przeprowadzenia skalowania wielowymiarowego, w którym niepodobieństwa między obiektami i między respondentami traktuje się jako brakujące dane.

Początki prac nad analizą *unfolding* sięgają 1950 r., kiedy to Coombs zaproponował jednowymiarowy model niemetryczny. Dziesięć lat później Bennett i Hays [1960] uogólnili model Coombsa, prezentując możliwości wykorzystania jego idei w przestrzeni wielowymiarowej. Kluczowy wkład dla rozwoju metody miały ponadto m.in. prace Heisera [1981], DeSarbo i Carrolla [1985], Kima, Rangaswamy'ego i DeSarbo [1999], Businga, Groenena i Heisera [2005].

3. Zdegenerowane rozwiązania w analizie *unfolding*

Przyjmijmy, że współrzędne punktów idealnych x_{ia}, gdzie $i = 1, \ldots, n$, $a = 1, \ldots, r$ (n – liczba punktów idealnych, r – wymiar przestrzeni, w której prezentowane są wyniki analizy), tworzą macierz **X**, natomiast y_{ja}, gdzie $j = 1, \ldots, m$ (m – liczba obiektów), są współrzędnymi punktów reprezentujących obiekty i tworzą macierz **Y**. Celem analizy *unfolding* jest znalezienie dla ustalonych ocen preferencji δ_{ij} takiego odwzorowania φ, dla którego:

$$d_{ij} \approx \hat{d}_{ij} = f(\delta_{ij}), \tag{1}$$

gdzie: $d_{ij} = \sqrt{\sum_{a=1}^{r} (x_{ia} - y_{ja})^2}$ – odległość między $\mathbf{x}_i$ a $\mathbf{y}_j$,

$\hat{d}_{ij}$ – monotoniczna funkcja regresji między d_{ij} a δ_{ij}.

Dla ocen preferencji mierzonych na skali porządkowej funkcja φ spełnia zależność [Takane, Young, de Leeuw 1977]:

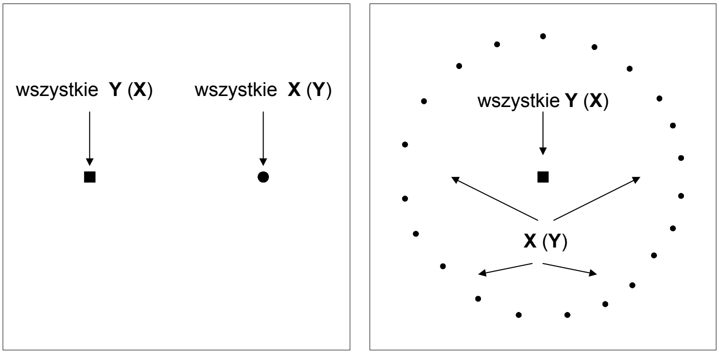
$$\delta_{ij} \prec \delta_{i'j'} \Rightarrow \hat{d}_{ij} \leq \hat{d}_{i'j'}. \tag{2}$$

W analizie *unfolding* współrzędne punktów **X** i **Y** wyznaczane są tak, aby minimalizowały wartość funkcji dopasowania STRESS[1]:

$$S = \sqrt{\frac{\sum_i \sum_j \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_i \sum_j d_{ij}^2}}. \tag{3}$$

W praktyce minimalna wartość funkcji dopasowania nie zawsze jednak może oznaczać idealne dopasowanie otrzymanej konfiguracji punktów do danych wejściowych. Często uzyskujemy wyniki analiz, w których wartość funkcji STRESS jest równa zeru, a struktura danych wejściowych jest nadal ukryta i nieinterpretowalna. W takim przypadku mówimy o rozwiązaniu zdegenerowanym. Rozwiązanie (**X**, **Y**) nazywamy całkowicie zdegenerowanym, jeżeli funkcja dopasowania jest równa zeru i dla każdego i, j $d_{ij} = d$ (d – pewna dodatnia stała). Tym samym wszystkie przekształ-

[1] Inne postacie funkcji dopasowania prezentują m.in. [Borg, Groenen 2005, s. 251-253; Zaborski 2001, s. 53].



Rys. 2. Zdegenerowane rozwiązania w analizie *unfolding*
Źródło: opracowano na podstawie pracy [Borg, Groenen 2005, s. 301].

cone odległości $\hat{d}_{ij}$ spełniające warunek (2) są równe d (zob. np. [Cox, Cox 2001, s. 65; Zaborski 2001, s. 79]) . Przyjmuje się jednocześnie, że nie wszystkie oceny preferencji są powiązane, czyli istnieją takie i, j oraz i', j', dla których $\delta_{ij} \neq \delta_{i'j'}$. Przykłady zdegenerowanych rozwiązań w analizie *unfolding* prezentuje rys. 2.

4. Charakterystyka modelu PREFSCAL

Model PREFSCAL jest rozwinięciem algorytmu PROXSCAL, który w kolejnych cyklach iteracyjnych przybliża optymalne rozwiązanie za pomocą metody majoryzacji funkcji dopasowania. Jest on dostępny w pakiecie statystycznym SPSS od wersji 14. W modelu PREFSCAL uniknięcie równych wartości $\hat{d}_{ij}$, a tym samym zdegenerowanego rozwiązania, osiągnięto przez modyfikację funkcji dopasowania STRESS.

W konstrukcji miary dopasowania Busing, Groenen i Haiser [2005] zaproponowali wykorzystanie współczynnika zmienności Pearsona, który definiują jako:

$$\nu(\hat{d}) = \frac{\text{odchylenie standardowe } (\hat{\mathbf{d}})}{\text{średnia } (\hat{\mathbf{d}})} = \frac{\sqrt{K^{-1}\sum_{k=1}^{K}\left(\hat{d}_k - \bar{\hat{d}}_k\right)^2}}{\bar{\hat{d}}}, \tag{4}$$

gdzie $\bar{\hat{d}} = K^{-1}\sum_{k=1}^{K}\hat{d}_k$ (K – liczba wszystkich $\hat{d}_{ij}$).

Współczynnik zmienności traktowany jest jako diagnostyka służąca identyfikacji rozwiązań o równych wartościach $\hat{d}_{ij}$. Model PREFSCAL wykorzystuje współczynnik zmienności do „karania” funkcji dopasowania w przypadku równych wartości $\hat{d}_{ij}$. W omawianym modelu funkcja dopasowania przyjmuje postać:

$$S_p(\hat{\mathbf{d}},\mathbf{X},\mathbf{Y}) = S^{2\lambda}(\hat{\mathbf{d}},\mathbf{X},\mathbf{Y})\left(1+\frac{\omega}{\nu^2(\hat{\mathbf{d}})}\right), \tag{5}$$

gdzie: $S(\hat{\mathbf{d}},\mathbf{X},\mathbf{Y})$ – znormalizowana funkcja STRESS zdefiniowana równaniem (3), $0<\lambda<1$, $\omega>0$ – parametry karania.

Parametr λ wpływa na równowagę pomiędzy znormalizowaną wartością funkcji STRESS a czynnikiem kary $\left(1+\frac{\omega}{\nu^2(\hat{\mathbf{d}})}\right)$. Im mniejsza wartość parametru λ, tym większy wpływ czynnika kary na wartość funkcji dopasowania. Parametr ω określa wielkość kary nałożonej na znormalizowaną wartość funkcji STRESS przy określonym poziomie $\nu^2(\hat{\mathbf{d}})$.

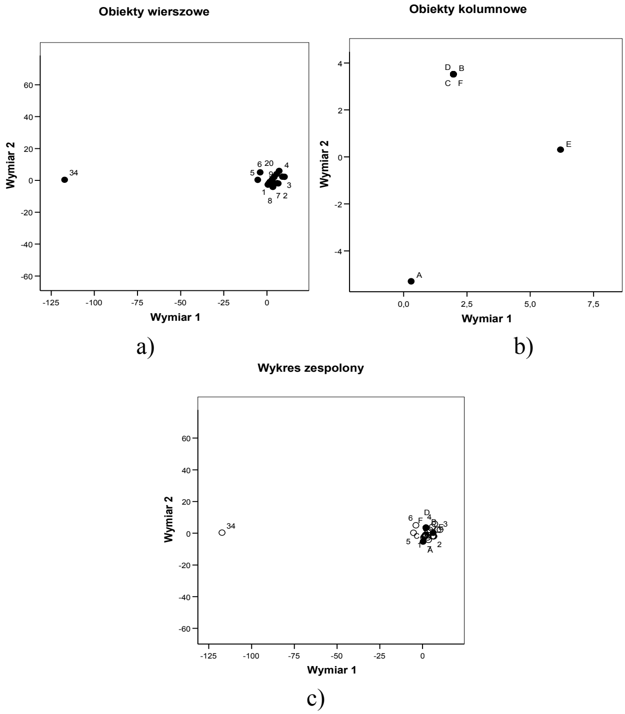
5. Przykład empiryczny

W celu identyfikacji preferencji 38 respondentom przedstawiono tytuły 9 rozrywkowych programów telewizyjnych (zob. [Zaborski 2002]). Następnie poproszono ich o określenie swoich preferencji poprzez nadanie poszczególnym programom rang, przy czym rangi były kolejnymi liczbami naturalnymi od 1 do 9. Liczba 1 oznaczała ocenę najlepszą, natomiast liczba 9 – ocenę najgorszą. Na podstawie otrzymanej w ten sposób macierzy preferencji przeprowadzono analizę *unfolding* przy wykorzystaniu algorytmu PREFSCAL, przy czym wartości parametrów karania ustalono następująco: $\lambda=1$ i $\omega=0$. Oznacza to, że konfigurację punktów wyznaczano przez minimalizację funkcji dopasowania STRESS bez czynnika kary. Wyniki analizy prezentuje rys. 3.

Rozkład punktów na rys. 3 wskazuje na zdegenerowane rozwiązanie. Prawie wszystkie punkty reprezentujące respondentów znajdują się w bardzo bliskim sąsiedztwie, a siedem z dziewięciu punktów reprezentujących badane programy ma takie same współrzędne. Trywialność rozwiązania potwierdzają również wartości miar zawarte w tab. 1. W omawianym przykładzie osiągnięto wprawdzie bardzo niską wartość STRESS-u, co mogłoby świadczyć o doskonałym dopasowaniu otrzymanej konfiguracji punktów do danych wejściowych, jednak indeks braku degeneracji Sheparda będący ilorazem liczby różnych odległości[2] i liczby wszystkich odległości jest równy tylko 0,158626. Również wartość współczynnika zmienności dla przekształconych odległości $\hat{d}_{ij}$ jest bardzo niska i wynosi 0,000456.

W celu znalezienia niezdegenerowanego rozwiązania podobną analizę przeprowadzono dla parametrów $\lambda=0,5$ i $\omega=1$. Dla takich wartości parametrów karania rozwiązanie wydaje się najlepsze spośród wariantów przedstawionych w tab. 1 (niska wartość STRESS przy jednoczesnej wysokiej wartości wskaźnika braku degeneracji). Wyniki analizy prezentuje rys. 4.

[2] Przez różne odległości rozumie się takie, których różnica jest większa niż 0,001 odległości przeciętnej.



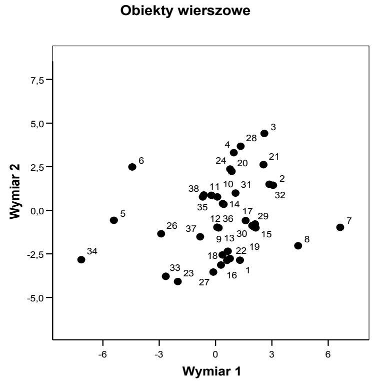
Rys. 3. Konfiguracje punktów dla $\lambda = 1$ i $\omega = 0$ reprezentujące: a) respondentów; b) obiekty; c) respondentów i obiekty

Źródło: opracowanie własne z wykorzystaniem pakietu *SPSS*.

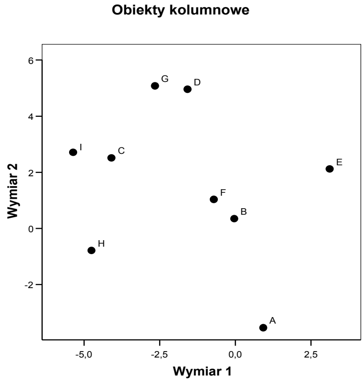
Tabela 1. Wybrane miary dopasowania dla różnych wartości parametrów karania

Siła kary		$\lambda = 0{,}1$		$\lambda = 0{,}5$		$\lambda = 1{,}0$	
Zakres działania kary		$\omega = 0{,}5$	$\omega = 1{,}0$	$\omega = 0{,}5$	$\omega = 1{,}0$	$\omega = 0{,}0$	$\omega = 0{,}5$
Końcowa wartość funkcji		1,049175	1,182228	0,433089	0,596575	0,001754	0,054401
Cząstkowe wartości funkcji	STRESS	0,324284	0,338814	0,079759	0,151843	0,001754	0,007615
	czynnik kary	4,983709	15,741716	2,351656	2,343880	1,000000	7,143845
Niedopasowanie	znormalizowany STRESS	0,101414	0,110690	0,006358	0,022801	0,000003	0,000058
	STRESS-I Kruskala	0,318456	0,332701	0,079737	0,150999	0,001754	0,007615
Dobroć dopasowania	Rho Spearmana	0,732961	0,732522	0,876279	0,877991	0,852073	0,843348
	Tau-b Kendalla	0,588627	0,587165	0,787397	0,791849	0,775032	0,766987
Współczynnik zmienności	zmienność przekształconych odległości	0,667774	0,677729	0,359893	0,504081	0,000456	0,075267
Wskaźnik degeneracji	indeks braku degeneracji Sheparda	0,777778	0,777047	0,547515	0,722953	0,158626	0,159357

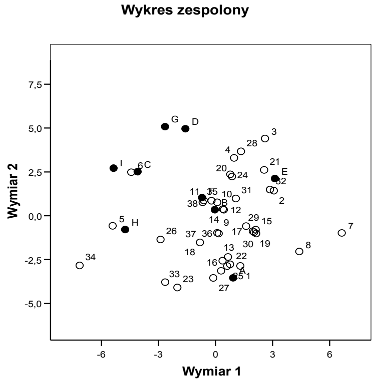
Źródło: opracowanie własne z wykorzystaniem pakietu *SPSS*.



a)



b)



c)

Rys. 4. Konfiguracje punktów dla $\lambda = 0{,}5$ i $\omega = 1{,}0$ reprezentujące: a) respondentów; b) obiekty; c) respondentów i obiekty

Źródło: opracowanie własne z wykorzystaniem pakietu *SPSS*.

Rysunek 4c wskazuje na programy najbardziej preferowane przez oceniających je respondentów. Są to: program A, posiadające tych samych zwolenników programy B i F oraz program E. Pozostałe programy, a zwłaszcza G i D, wyraźnie oddalone są od skupisk punktów idealnych.

6. Podsumowanie

Pojawianie się równych przekształconych odległości, a tym samym trywialnych rozwiązań, jest częstym zjawiskiem w wynikach skalowania wielowymiarowego. Wprawdzie niektórzy autorzy wskazują na możliwe przyczyny degeneracji (zob. np. [Kim, Rangaswamy, DeSarbo 1999; Busing, Groenen, Heiser 2005]), jednak nie są one w pełni wyjaśnione. Zastosowanie współczynnika zmienności w konstrukcji funkcji dopasowania modelu PREFSCAL sprawia, że w kolejnych cyklach iteracyjnych wyznaczane są konfiguracje o możliwie różnych odległościach między punktami. Z jednej strony pozwala to na uniknięcie zdegenerowanego rozwiązania, a jeżeli wartość zmodyfikowanej funkcji STRESS przyjmuje wartość zero, świadczy to o doskonałym dopasowaniu konfiguracji punktów do danych wejściowych.

Literatura

Bennett J.F., Hays W.L. (1960), *Multidimensional unfolding: determining the dimensionality of ranked preference data*, „ Psychometrika” no 25, s. 27-43.

Borg I., Groenen P. (2005), *Modern multidimensional scaling. Theory and applications*. Second Edition, Springer-Verlag, New York.

Busing F.M.T.A., Groenen P.J.K., Heiser W.J. (2005), *Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation*, „Psychometrika” no 1, s. 71-79.

Coombs D.V. (1950), *Psychological scaling without a unit of measurement*, „Psychological Review” no 57, s. 145-158.

Cox T.F., Cox M.A.A. (2001), *Multidimensional scaling. Second edition*, Chapman and Hall, Londyn.

DeSarbo W.S., Carroll J.D. (1985), *Three-way metric unfolding via alternating weigted least squares*, „Psychometrika” no 50, s. 275-300.

Heiser W.J. (1981), *Unfolding analysis of proximity data*, Leiden.

Kim C., Rangaswamy A., DeSarbo W.S. (1999), *A quasi-metric approach to multidimensional unfolding for reducing the occurence of degenerate solutions*, „Multivariate Behavioral Research” no 34, s. 143-180.

Takane Y., Young F.W., de Leeuw J. (1977), *Nonmetric individualdifferences MDS: An alternating least squares method with optimal scaling features*, „Psychometrika” no 42, s. 7-67.

Zaborski A. (2001), *Skalowanie wielowymiarowe w badaniach marketingowych*, AE, Wrocław.

Zaborski A. (2002), *Unfolding jako model pomiaru preferencji w skalowaniu wielowymiarowym*, [w:] Taksonomia 9, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 942, AE, Wrocław, s. 128-137.

POSSIBILITIES OF AVOIDING DEGENERATE SOLUTIONS IN UNFOLDING ANALYSIS BY USING PREFSCAL ALGORITHM

Summary

The article presents PREFSCAL algorithm in which avoiding degeneracy in multidimensional unfolding is possible by modification of the loss function. In the structure of STRESS function, the variation coefficient is used as a diagnostic for identifying solutions with constant interpoint distances. Finally, empirical data set was analyzed by PREFSCAL method to show the benefits of the modification of the loss function.