Fragment rozdziału (wersja robocza 1): Chybalski F., *Analizy wielowymiarowe*, [w:] I. Staniec (red.) „Metody ilościowe w zarządzaniu organizacją", C.H. Beck, Warszawa 2013, s. 86-105.

5 Analizy wielowymiarowe

Filip Chybalski

Opis problemu

Grupa finansowa PF opracowuje swoją strategię wejścia na rynek europejski z produktami finansowymi dla osób starszych, szczególnie dla emerytów. Sztandarowym produktem PF skierowanym do tej grupy wiekowej jest odwrócona hipoteka, polegająca na tym, że instytucja finansowa wypłaca beneficjentowi ustaloną rentę dożywotnią, a w zamian po jego śmierci przejmuje nieruchomość, w której on zamieszkiwał.

W ramach analizy rynku, PF zamierza przeprowadzić badanie porównawcze, w którym porówna sytuację materialną osób starszych w krajach europejskich. Na podstawie sformułowanych wniosków, dział marketingu opracuje odpowiednie strategie reklamowe i akwizycyjne dla poszczególnych krajów. Zarząd PF postawił następujące zadanie przed pracownikami działu analitycznego firmy: przygotowanie zestawienia krajów europejskich pod względem sytuacji materialnej w populacji osób starszych.

Opis wykorzystanej metody: wielowymiarowa analiza porównawcza (WAP)

Sytuacji materialnej czy też dochodowej określonej populacji nie można zmierzyć za pomocą jednego, prostego wskaźnika, ponieważ charakteryzuje ją wiele cech statystycznych, takich jak np. poziom dochodów, zróżnicowanie dochodów, ubóstwo. Mamy zatem do czynienia z tzw. zjawiskiem złożonym, tzn. opisanym za pomocą liczby zmiennych większej od 1. Dlatego też w celu zrealizowania zadań postawionych przez zarząd PF, dział analityczny skorzysta z wielowymiarowej analizy porównawczej (WAP), która umożliwia porównywanie między sobą wielu obiektów, opisanych za pomocą wielu cech (zmiennych, wskaźników, mierników). Cechy te mogą mieć charakter:

- stymulant, w przypadku których korzystne są wyższe wartości zmiennych,
- destymulant, w przypadku których korzystne są niższe wartości zmiennych,
- nominant, w przypadku których pożądane są określone wartości normatywne zmiennych (konkretna wartość zmiennej lub przedział wartości).

Pierwszym krokiem w wykorzystaniu WAP do analizy określonego zjawiska złożonego jest dobór zmiennych, które **powinny spełniać określone postulaty**: merytoryczne, formalne oraz weryfikowalne statystycznie.

Postulaty merytoryczne badacz uwzględnia na etapie doboru potencjalnych zmiennych diagnostycznych. Bardzo ważna jest w tym miejscu znajomość badanego zjawiska, bo właśnie w oparciu o nią badacz decyduje, które zmienne powinny być uwzględniane w analizie, a które nie.

Postulatami o charakterze formalnym są przede wszystkim dostępność danych oraz ich porównywalność. **Porównywalność danych**, szczególnie istotną w analizach międzynarodowych, należy rozpatrywać w trzech aspektach i na tej podstawie można wyróżnić:[1]

- porównywalność konceptualną, zgodnie z którą pomiary powinny odnosić się do tych samych pozycji lub pojęć,
- porównywalność statystyczną oznaczająca, że dla wszystkich pozycji powinny być zastosowane metody zbierania danych, akceptowane w badaniach statystycznych,
- porównywalność interpretacyjną, zgodnie z którą badane kategorie powinny być interpretowane w ten sam sposób we wszystkich badanych krajach z uwzględnieniem występujących w nich uwarunkowań.

W ostatnim etapie doboru zmiennych należy uwzględnić **postulaty statystyczne** pod adresem zmiennych, ponieważ powinny się one charakteryzować:

- **odpowiednio wysoką zmiennością**, co świadczy o ich zdolnościach dyskryminujących badane obiekty,
- **odpowiednio dużą pojemnością informacyjną**, co oznacza, że w zbiorze zmiennych diagnostycznych powinny znaleźć się te zmienne, które zawierają największy ładunek informacyjny. W uproszczeniu można powiedzieć, że spośród dwóch zmiennych statystycznie podobnych (co oznacza, że ich wykresy mają podobny kształt), do ostatecznego zbioru zmiennych diagnostycznych powinna trafić ta, która jest bogatsza w informację.

Zmienne zawarte w zbiorze potencjalnych zmiennych diagnostycznych można zbadać pod względem spełniania powyższych kryteriów za pomocą niżej opisanych narzędzi statystycznych. Do oceny zmienności można wykorzystać miary klasyczne lub pozycyjne. W analizowanym przykładzie zostanie wykorzystana miara klasyczna, jaką jest **współczynnik zmienności**, wyrażony wzorem:

$$v_x^K = \frac{s_x}{\bar{x}} \tag{5.1}$$

[1] Panek T., Szulc A. (red.), *Statystyka społeczna. Wybrane zagadnienia*, Szkoła Główna Handlowa, Warszawa 2006, s. 24

gdzie:

$\bar{x}$ - średnia arytmetyczna wartość zmiennej *X*, szacowana według wzoru:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \tag{5.2}$$

s_x – odchylenie standardowe cechy *X*:

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{5.3}$$

Przyjmując określoną wartość progową współczynnika zmienności, ze zbioru potencjalnych wskaźników diagnostycznych eliminuje się te, w przypadku których współczynnik zmienności jest mniejszy bądź równy założonej wartości progowej. Należy jednak pamiętać, że usuwając cechy charakteryzujące się małą zmiennością co prawda zwiększamy zdolności dyskryminacyjne w zbiorze zmiennych diagnostycznych, ale jednocześnie, poniekąd w sztuczny sposób, zmniejszamy podobieństwo obiektów. Przykładowo, jeżeli w zbiorze potencjalnych zmiennych diagnostycznych jest 10 zmiennych, dobranych przez badacza w oparciu o kryteria merytoryczne, spośród których 7 charakteryzuje się małą zmiennością (czyli mają mało zróżnicowane wartości), ich usunięcie powoduje wyeliminowanie tego podobieństwa obiektów. Wówczas wyniki analizy wielowymiarowej mogą sugerować małe podobieństwo obiektów, podczas gdy w oparciu o wszystkie zmienne, które badacz uznał za ważne, to podobieństwo byłoby ocenione jako większe. W ten sposób kryteria statystyczne mogą w mniejszym lub większym stopniu zniekształcić i wypaczyć wyniki analizy i ważne jest, by badacz miał tego świadomość.

Kolejną istotną kwestią jest **ocena potencjału informacyjnego każdej zmiennej**, która powinna być jak najsłabiej skorelowana z pozostałymi zmiennymi diagnostycznymi i jak najsilniej skorelowana ze zmiennymi, które na etapie doboru cech diagnostycznych zostały wyeliminowane. Oznacza to bowiem, że dana zmienna wchodząc do zbioru zmiennych diagnostycznych dobrze reprezentuje zmienne wcześniej usunięte. Metody doboru zmiennych diagnostycznych ze względu na ich potencjał informacyjny bazują na **współczynniku korelacji**, wyrażonym wzorem:

$$r_{XY} = \frac{\operatorname{cov}\left(X,Y\right)}{s_x s_Y}, \tag{5.4}$$

gdzie:

cov (*X*,*Y*) – kowariancja cech *X* i *Y*, wyrażona wzorem:

$$\operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right). \tag{5.5}$$

W oparciu o obliczone dla wszystkich par potencjalnych zmiennych diagnostycznych współczynniki korelacji, tworzy się macierz korelacji o wymiarach *m* x *m* (*m* – liczba zmiennych), której element na przecięciu *i*-tego wiersza i *j*-tej kolumny jest współczynnikiem korelacji między *i*-tą a *j*-tą zmienną. Macierz ta jest oczywiście macierzą symetryczną.

Wśród najczęściej stosowanych metod doboru zmiennych diagnostycznych w oparciu o współczynniki korelacji można wymienić metodę parametryczną oraz metodą odwróconej macierzy. **Metoda parametryczna** ma dwie istotne wady, wskazywane w literaturze. Mianowicie jest ona wrażliwa na wartości odstające oraz uwzględnia wyłącznie bezpośrednie powiązania danej zmiennej z innymi zmiennymi, pomijając w ogóle powiązania pośrednie. **Metoda odwróconej macierzy korelacji** ma tę przewagę nad pierwszą, że jest wolna od drugiej ze wspomnianych wad, czyli uwzględnia także pośrednie relacje między zmiennymi.[2] Tę metodę zastosujemy przy doborze zmiennych diagnostycznych w naszym przypadku. Poniżej dokonano jej szczegółowej charakterystyki.

[2] Porównaj pozycje: Panek T., Statystyczne metody wielowymiarowej analizy porównawczej, Szkoła Główna Handlowa, Warszawa 2009, s. 22, Młodak A., Analiza taksonomiczna w statystyce regionalnej, Difin, Warszawa 2006, s. 31.