Anna Czopek

ANALIZA PORÓWNAWCZA EFEKTYWNOŚCI METOD REDUKCJI ZMIENNYCH – ANALIZA SKŁADOWYCH GŁÓWNYCH I ANALIZA CZYNNIKOWA

Wprowadzenie

Analiza składowych głównych i analiza czynnikowa to dwie najbardziej popularne metody pozwalające na sprowadzenie dużej liczby badanych zmiennych do znacznie mniejszej liczby wzajemnie niezależnych składowych głównych lub czynników. Nowe zmienne (składowe główne lub czynniki) zachowują stosunkowo dużą część informacji zawartych w zmiennych pierwotnych, a jednocześnie każda z nich jest nośnikiem innych treści merytorycznych. Obie powyższe metody redukcji zmiennych są często stosowane, gdyż zbyt duża ilość rozpatrywanych cech powoduje wzrost skali trudności interpretacji.

Zasadniczą przyczyną podjęcia tematu jest próba pokazania, że wyżej wymienionych metod, choć są bardzo podobne, nie można utożsamiać. Mimo tego, iż w obu przypadkach są obliczane wartości własne, ładunki czynnikowe itp., to jednak istnieją między nimi różnice w sposobie działania, o czym należy pamiętać. Zatem stosowanie tych nazw zamiennie jest niedopuszczalne.

Artykuł składa się z trzech części. Rozdziały pierwszy i drugi są poświęcone, odpowiednio, analizie składowych głównych i analizie czynnikowej, gdzie została dokonana krótka charakterystyka tych metod. W rozdziale trzecim, na podstawie przykładu empirycznego, porównano efektywność analizy składowych głównych i analizy czynnikowej.

1. Analiza składowych głównych

Początki techniki analizy składowych głównych pochodzą od Pearsona (1901). Jednak główny rozwój tej metody zawdzięcza się pracom amerykańskiego statystyka Hotellinga (1933), który wykorzystał ją do analizy testów osiągnięć szkolnych.

Podstawową ideą metody jest transformacja wyjściowego zbioru zmiennych $X_1, \dots, X_p$ na nowy zbiór zmiennych $Z_1, \dots, Z_p$, zwanych składowymi głównymi. W konsekwencji liczba głównych składowych jest równa liczbie zmiennych pierwotnych. W praktyce nie ma to jednak dużego znaczenia, gdyż liczbę składowych głównych ogranicza się w dalszych rozważaniach do kilku najważniejszych. Zatem celem analizy składowych głównych jest redukcja liczby zmiennych przy zachowaniu tak dużej zmienności danych, jak to tylko możliwe.

Model matematyczny w analizie składowych głównych jest sformułowany w postaci następującego układu równań liniowych:

$$\begin{aligned} X_1 &= a_{11}Z_1 + a_{12}Z_2 + \cdots + a_{1p}Z_p \\ X_2 &= a_{21}Z_1 + a_{22}Z_2 + \cdots + a_{2p}Z_p \\ &\vdots \\ X_p &= a_{p1}Z_1 + a_{p2}Z_2 + \cdots + a_{pp}Z_p \end{aligned}$$

Zmienne rzeczywiste podlegające obserwacji X_i dla $i \in \{1, \dots, p\}$ są wyrażone jako kombinacje liniowe zmiennych nieobserwowalnych Z_j dla $j \in \{1, \dots, p\}$, zwanych składowymi głównymi. Współczynniki a_{ij} dla $i, j \in \{1, \dots, p\}$ określają wagę danej składowej w opisie zmiennych empirycznych.

1.1. Algorytm postępowania w analizie składowych głównych

Poniższe kroki opisują schemat postępowania w analizie składowych głównych [2; 5; 8].

Krok I – Sprawdzenie założeń

Przed rozpoczęciem analizy składowych głównych należy sprawdzić podstawowe założenie, aby ocenić zasadność jej zastosowania, a mianowicie skorelowanie zmiennych – im wyższe korelacje między zmiennymi pierwotnymi, tym bardziej uzasadnione jest wykorzystanie tej analizy. Korelację bada się analizując macierz korelacji dla zmiennych wziętych do analizy lub wykorzystując test Bartletta [8].

Należy również zwrócić uwagę na poniższe warunki [3; 8]:

1. **Normalność rozkładu** – założenie to nie jest konieczne, gdy analizuje się duży zbiór danych.
2. **Liczebność i reprezentatywność próby** – do analizy przystępuje się, gdy próba liczy co najmniej 50 obserwacji. Próbę należy pobrać w sposób losowy. Zbiór obserwacji musi być jednorodny.

3. **Punkty odstające** – punkty odstające niestety często zniekształcają prawdziwe zależności między zmiennymi. Dobrze jest na początku analizy wykryć takie punkty i usunąć je z danych.
4. **Braki danych** – w przypadku brakujących danych w analizowanej próbie należy zastąpić braki przez średnie lub usunąć przypadki z brakującymi danymi.

Krok II – Wybór odpowiedniej macierzy

Następnie należy przyjrzeć się początkowym zmiennym. Jeżeli analizowane zmienne są porównywalne (wyrażają się w tych samych jednostkach i są tego samego rzędu), to w dalszej analizie wykorzystuje się macierz kowariancji. Jeżeli natomiast zmienne mają różne jednostki lub są różnego rzędu, analizę składowych głównych przeprowadza się wykorzystując macierz korelacji. Jest to ważny krok rozpoczynający całą analizę, gdyż składowe główne otrzymane dla macierzy kowariancji i korelacji nie muszą być takie same.

Krok III – Wyznaczenie składowych głównych

Niech $\boldsymbol{X} = \left(X_1, \dots, X_p\right)^T$ będzie wektorem zmiennych wziętych do analizy. Składowe główne są kombinacją liniową zmiennych początkowych:

$$\begin{aligned} Z_1 &= a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p \\ Z_2 &= a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p \\ &\vdots \\ Z_p &= a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p \end{aligned}$$

Kolejnym krokiem jest wyznaczenie macierzy współczynników a_{ij} dla $i, j \in \{1, \dots, p\}$ dla zadanego z góry wektora obserwacji $\boldsymbol{X}$. Algorytm wyznaczania współczynników a_{ij} dla $i, j \in \{1, \dots, p\}$ bardzo dokładnie opisuje D.F. Morrison [5] i A. Stanisz [8].

Krok IV – Redukcja wymiaru – kryteria wyboru

Ważną informacją jest to, że każda kolejna wyznaczona składowa główna wyjaśnia coraz mniejszą część zmienności początkowych zmiennych. W jakimś momencie okaże się, że któraś z kolei składowa określa znikomą część zmienności. Należy zatem dokonać redukcji składowych, stosując w dalszych rozważaniach tylko najważniejsze.

Popularne kryteria redukcji [2; 4; 8]:

1. **Kryterium wystarczającej proporcji** – stopień wyjaśnionej wariancji oryginalnych zmiennych musi wynosić co najmniej 75%. W praktyce najczęściej już przy 2-3 głównych składowych stopień wyjaśnienia wariancji jest wystarczający.

2. **Kryterium Kaisera** – eliminacja składowych głównych, których wartości własne są mniejsze od 1.
3. **Wykres osypiska** – wyznaczenie na wykresie liniowym kolejnych wartości własnych. Interpretacja polega na znalezieniu miejsca, od którego na prawo występuje łagodny spadek wartości własnych. Nie powinno się uwzględniać więcej czynników, niż te znajdujące się po lewej stronie tego punktu.

Wybór odpowiedniego kryterium leży w gestii statystyka, dlatego też decyzja ta jest dosyć subiektywna i wpływa na rezultaty analizy.

Krok V – Interpretacja

Interpretację otrzymanych wyników przeprowadza się za pomocą tzw. ładunków czynnikowych. Ładunki czynnikowe są współczynnikami korelacji pomiędzy daną zmienną a składowymi.

Jeżeli powyższa analiza jest przeprowadzana na podstawie macierzy kowariancji, to współczynnik korelacji pomiędzy i-tą zmienną X_i a j-tą składową Z_j dla $i, j \in \{1, \dots, p\}$ oblicza się ze wzoru:

$$r_{X_i,Z_j} = \frac{cov(X_i, Z_j)}{s_i\sqrt{\lambda_j}} = \frac{\lambda_j a_{ij}}{s_i\sqrt{\lambda_j}} = \frac{\sqrt{\lambda_j}a_{ij}}{s_i}$$

gdzie:
s_i – odchylenie standardowe zmiennej X_i,
λ_j – wariancja składowej głównej Z_j, a także j-ta co do wielkości wartość własna macierzy korelacji (kowariancji), na której opiera się cała analiza,
$\sqrt{\lambda_j}$ – odchylenie standardowe składowej Z_j.

Jeśli natomiast składowe są generowane z macierzy korelacji, to:

$$r_{X_i,Z_j} = \sqrt{\lambda_j}a_{ij}$$

Suma wszystkich wartości własnych macierzy korelacji (kowariancji) $\lambda_1 + \cdots + \lambda_p$ jest całkowitą wariancją układu. Dzięki temu można zdefiniować część całkowitej wariancji wyznaczoną przez j-tą składową:

$$h_j = \frac{\lambda_j}{\lambda_1 + \cdots + \lambda_p} \cdot 100\%$$

Natomiast procentowy udział zmienności całkowitej wyjaśnionej przez k pierwszych składowych oblicza się następująco:

$$H_k = \sum_{j=1}^{p} h_j$$

2. Analiza czynnikowa

Twórcami głównej koncepcji tej metody są psychologowie CH. Spearman (1904) i L.L. Thurstone (1913). Ch. Spearman wprowadził pojęcie pojedynczego czynnika ogólnego dla wyjaśnienia wyników testów inteligencji. Dopiero L.L. Thurstone stworzył podstawy teoretyczne analizy czynnikowej. Celem analizy czynnikowej jest dążenie do wyodrębnienia wszystkich czynników, które mogą rzeczywiście tkwić w korelacjach danego układu zmiennych, jednocześnie zachowując jak najwięcej informacji zawartych w zmiennych pierwotnych, a następnie redukcja tych czynników.

Model analizy czynnikowej konstruuje się jako założenie wstępne, które jest sformułowane w postaci układu równań:

$$\begin{aligned} X_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + b_1U_1 \\ X_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + b_2U_2 \\ &\vdots \\ X_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + b_pU_p \end{aligned}$$

gdzie $m < p$.

Standaryzowane zmienne pierwotne X_i dla $i \in \{1, \dots, p\}$ są wyrażone jako liniowe funkcje zmiennych nieobserwowalnych F_j dla $j \in \{1, \dots, m\}$, zwanych czynnikami wspólnymi i pojedynczego czynnika losowego U_i dla $i \in \{1, \dots, p\}$, zwanego czynnikiem swoistym. Współczynniki a_{ij} oraz b_i dla $i \in \{1, \dots, p\}$, $j \in \{1, \dots, m\}$ są zwane ładunkami czynnikowymi i określają wagę danego czynnika w opisie zmiennych empirycznych.

W analizie czynnikowej przyjmuje się dwa założenia o zmiennych i czynnikach:

1. Zmienne i czynniki są zestandaryzowane.
2. Czynniki wspólne F_j są ze sobą nieskorelowane, czynniki swoiste U_i są ze sobą nieskorelowane, czynniki wspólne F_j są nieskorelowane z czynnikami swoistymi U_i dla $i \in \{1, \dots, p\}$, $j \in \{1, \dots, m\}$.

2.1. Algorytm postępowania w analizie czynnikowej

Poniższe kroki opisują schemat postępowania w analizie czynnikowej [2; 5; 8; 9].

Krok I – Sprawdzenie założeń

Założenia w analizie czynnikowej są podobne jak w analizie składowych głównych z tym wyjątkiem, że zmienne pierwotne powinny mieć rozkład normalny lub być doprowadzone do takiej postaci drogą odpowiednich transformacji. Punktem wyjścia obliczeń jest macierz korelacji. Należy dokonać wstępnej oceny istniejących korelacji.

Krok II – Metody estymacji modelu analizy czynnikowej

Rozwiązanie analizy czynnikowej polega na wyznaczeniu układu czynników wspólnych F_j dla $j \in \{1, \dots, m\}$, co jest równoważne z określeniem dla każdego czynnika F_j odpowiadającego mu wektora $(a_{1j}, \dots, a_{pj})$. Dokonuje się tego wykorzystując jedną z podstawowych metod estymacji, do których m.in. należą [1; 5; 7; 8; 9]:

1. **Metoda głównych składowych** – opracowana przez Hotellinga (1933).
2. **Metoda głównego czynnika** – opracowana przez Harmana (1960).
3. **Metoda największej wiarygodności** – opracowana przez Lawleya (1940).
4. **Metoda centroidalna** – opracowana przez Thurstone'a (1931).

Największe uznanie matematyków zdobyła metoda głównych składowych. Nie bez przyczyny jest ona ustawiona jako metoda domyślna w programie Statistica w analizie czynnikowej. Wybór każdej z tych metod jest zawsze obciążony mniejszą czy większą dozą arbitralności.

Krok III – Redukcja wymiaru – kryteria wyboru

Kryteria redukcji liczby czynników są analogiczne jak w analizie składowych głównych. Natomiast opierając analizę czynnikową na metodzie największej wiarygodności, można za pomocą istniejącego testu dobroci dopasowania określić, czy ilość wybranych czynników jest właściwą liczbą dla danego modelu, czy też nie [2; 5; 9].

Krok IV – Rotacja czynników

Często zdarza się, że zmienna ma wysokie ładunki na kilku czynnikach, co uniemożliwia jednoznaczną interpretację. W takiej sytuacji należy przeprowadzić rotację czynników. W większości przypadków rotacja czynników redukuje dwuznaczność interpretacji, jaka może wystąpić w rozwiązaniu bez rotacji.

Dzięki obrotowi można łatwiej utożsamić każdy czynnik ze zmiennymi, z którymi jest mocno skorelowany.

Ustalenie najwłaściwszej pozycji układu odniesienia jest jednym z najtrudniejszych kroków. Według L.L. Thurstone'a (1935) należy dążyć do tzw. **prostej struktury**, która znacznie ułatwia interpretację wyników. Prostota takiej struktury ładunków czynnikowych polega na tym, że każda zmienna ma stosunkowo najprostszą zawartość czynnikową, tj. dominujący ładunek jakiegoś jednego czynnika i odwrotnie – miarą danego czynnika są tylko niektóre spośród analizowanych zmiennych. W praktyce rzadko można doprowadzić do struktury czynnikowej spełniającej kryteria struktury prostej, należy jednak dążyć do uzyskania wyniku najbardziej do niej zbliżonego.

Do wykonania rotacji najczęściej stosuje się metodę VARIMAX lub QUARTIMAX [5; 9], które ostatecznie decydują o interpretacji modelu, gdyż różne metody dają różne pozycje układów osi czynników.

Krok V – Interpretacja

Podstawowym zadaniem analizy czynnikowej jest wyznaczenie macierzy współczynników a_{ij} zwanych ładunkami czynnikowymi dla $i \in \{1, \dots, p\}$, $j \in \{1, \dots, m\}$. Ładunki te można interpretować w ten sposób, że waga czynnika jest współczynnikiem korelacji między zmienną a czynnikiem. Zatem:

$$r_{X_i F_j} = a_{ij}$$

$$\text{dla } i \in \{1, \dots, p\}, j \in \{1, \dots, m\}.$$

Do interpretacji otrzymanych wyników szuka się tych zmiennych, które mają najwyższe (w wartościach bezwzględnych) wartości ładunków czynnikowych dla danych czynników. Ładunki czynnikowe opisują wkład zmiennej do poszczególnych czynników.

Część całkowitej wariancji wyjaśnionej przez j-ty czynnik jest obliczany ze wzoru:

$$h_j = \frac{\lambda_j}{\lambda_1 + \dots + \lambda_m} \cdot 100\%$$

gdzie:
λ_j – j-ta wartość własna macierzy korelacji dla $j \in \{1, \dots, m\}$.

Natomiast procentowy udział zmienności całkowitej wyjaśnionej przez k pierwszych czynników oblicza się następująco:

$$H_k = \sum_{j=1}^{m} h_j$$

3. Porównanie efektywności analizy składowych głównych i analizy czynnikowej

3.1. Informacje wstępne

Do badań posłużyły dane z Rocznika Statystycznego Pracy 2010. Analizie poddano 311 powiatów Polski ze względu na osiem zmiennych:

1. Bezrobotni poprzednio pracujący – BPP.
2. Bezrobotni zwolnieni z przyczyn dotyczących zakładów pracy – BZ.
3. Bezrobotni zamieszkali na wsi – BZW.
4. Bezrobotni nieposiadający prawa do zasiłku – BNPZ.
5. Zatrudnieni w warunkach zagrożenia związanego ze środowiskiem pracy – ZŚP.
6. Zatrudnieni w warunkach zagrożenia związanego z uciążliwością pracy – ZUP.
7. Zatrudnieni w warunkach zagrożenia związanego z czynnikami mechanicznymi – ZCM.
8. Poszkodowani w wypadku przy pracy – PPP.

3.2. Wyniki analizy empirycznej

W artykule tym dokonano redukcji liczby zmiennych opisujących zróżnicowanie powiatów Polski. Uzyskane wyniki pozwalają na porównanie metody analizy składowych głównych oraz analizy czynnikowej, wskazując przede wszystkim stopień efektywności każdej z nich. Obie analizy zostały przeprowadzone za pomocą programu Statistica.

Przeprowadzając badanie za pomocą analizy składowych głównych, wykorzystano macierz korelacji i otrzymano następujące rezultaty:

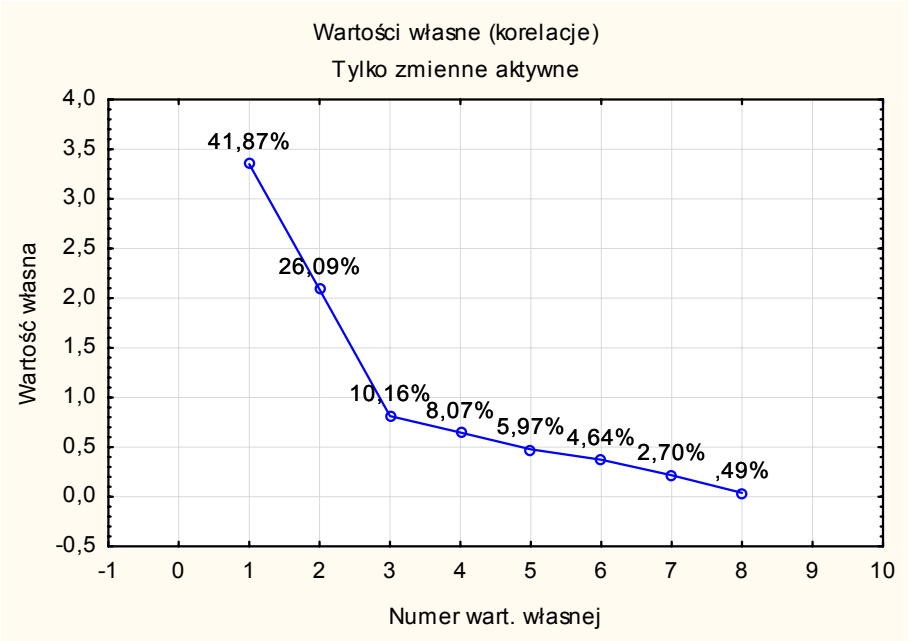
1. Korzystając z kryterium wystarczającej proporcji, dokonano redukcji ośmiu zmiennych do trzech składowych głównych. Na podstawie tabeli 1 można odczytać, iż pierwsza składowa wyjaśnia niecałe 42% całkowitej zmienności. Druga składowa wyjaśnia 26%, a trzecia 10%, co daje łącznie ponad 78% całkowitej zmienności.

Tabela 1

Wartości własne wyznaczone dla analizowanych danych, procent całkowitej wariancji wyjaśnionej przez j-tą składową (h_j), skumulowane wartości własne oraz skumulowany procent wyjaśnionej wariancji (H_k)

	Wartość własna	h_j (%)	Skumulowana wartość własna	H_k (%)
Z_1	3,349941	41,87426	3,34994	41,87426
Z_2	2,087453	26,09316	5,43739	67,96743
Z_3	0,813020	10,16275	6,25041	78,13018
Z_4	0,645555	8,06943	6,89597	86,19961
Z_5	0,477604	5,97005	7,37357	92,16966
Z_6	0,371222	4,64027	7,74479	96,80993
Z_7	0,215750	2,69689	7,96055	99,50681
Z_8	0,039455	0,49319	8	100

Wykres osypiska poniżej potwierdza tę decyzję (rysunek 1).



Rys. 1. Wykres osypiska

Źródło: Opracowanie własne z wykorzystaniem programu Statistica.

2. Ładunki czynnikowe traktuje się jako korelacje między zmiennymi a składowymi. Dla trzech pierwszych składowych ładunki przedstawiono w tabeli 2.

Tabela 2

Ładunki czynnikowe dla trzech pierwszych składowych

	Z_1	Z_2	Z_3
BPP	**-0,833739**	0,449623	0,073746
BZ	-0,511962	0,137246	**-0,818869**
BZW	-0,564090	**0,673689**	0,213194
BNPZ	**-0,826374**	0,505603	0,112882
ZŚP	-0,549488	**-0,655135**	-0,003492
ZUP	-0,534307	-0,460750	0,279763
ZCM	-0,539138	-0,511579	-0,020974
PPP	**-0,716612**	-0,503744	0,011093

Pierwsza składowa ma najwyższe, ujemnie ładunki czynnikowe ze zmiennymi BPP, BNPZ, PPP. Określa ona zatem bezrobotnych poprzednio pracujących, bezrobotnych nieposiadających prawa do zasiłku, poszkodowanych w wypadku przy pracy. Druga składowa ma najwyższe ładunki ze zmiennymi BZW i ZŚP, lecz korelacje te nie są zbyt wysokie. Obie zmienne oddziałują w sposób przeciwny na tą składową, BZW dodatnio, a ZŚP ujemnie. Trzecia składowa najsilniej i ujemnie jest związana ze zmienną BZ. Brakuje natomiast składowej najmocniej skorelowanej ze zmiennymi ZUP i ZCM. Obie te zmienne mają podobne (w wartościach bezwzględnych) wartości ładunków dla dwóch składowych – pierwszej i drugiej. Opisana struktura jest daleka od spełnienia warunków tzw. prostej struktury.

Warto sprawdzić, czy dodanie czwartej składowej głównej nie poprawi powyższej sytuacji. Jest to dość ryzykowne posunięcie, znacznie wpływające na rezultaty, gdyż jedynie kryterium wystarczającej proporcji jest spełnione. Wyniki przedstawia tabela 3.

Tabela 3

Ładunki czynnikowe dla czterech pierwszych składowych

	Z_1	Z_2	Z_3	Z_4
1	2	3	4	5
BPP	**-0,833739**	0,449623	0,073746	-0,012254
BZ	-0,511962	0,137246	**-0,818869**	0,179866
BZW	-0,564090	**0,673689**	0,213194	-0,086882
BNPZ	**-0,826374**	0,505603	0,112882	-0,013554

cd. tabeli 3

1	2	3	4	5
ZŚP	-0,549488	**-0,655135**	-0,003492	-0,028257
ZUP	-0,534307	-0,460750	0,279763	**0,561679**
ZCM	-0,539138	-0,511579	-0,020974	**-0,537121**
PPP	**-0,716612**	-0,503744	0,011093	-0,023245

Dodana czwarta składowa główna faktycznie jest w największym stopniu związana ze zmiennymi ZUP i ZCM, ale mimo wszystko w niezbyt wysokim. Krok ten jeszcze bardziej oddalił od spełnienia warunków prostej struktury, gdyż zmienne ZUP i ZCM mają teraz podobne (w wartościach bezwzględnych) wartości ładunków dla trzech składowych.

Przeprowadzając badanie za pomocą analizy czynnikowej, do wyodrębnienia czynników wykorzystano cztery metody: głównych składowych, głównego czynnika, największej wiarygodności oraz centroidalną. Za każdym razem wyniki są poprawione za pomocą rotacji Varimax. Wnioski przedstawiono poniżej:

1. Poniższe tabele zawierają: wartości własne wyznaczone dla analizowanych danych, procent całkowitej wariancji wyjaśnionej przez j-tą składową (h_j), skumulowane wartości własne oraz skumulowany procent wyjaśnionej wariancji (H_k) wyliczone za pomocą wymienionych wyżej metod.

Tabela 4

Metoda składowych głównych

	Wartość własna	h_j (%)	Skumulowana wartość własna	H_k (%)
F_1	3,349941	41,87426	3,34994	41,87426
F_2	2,087453	26,09316	5,43739	67,96743
F_3	0,813020	10,16275	6,25041	78,13018
F_4	0,645555	8,06943	6,89597	86,19961

Tabela 5

Metoda głównego czynnika

	Wartość własna	h_j (%)	Skumulowana wartość własna	H_k (%)
F_1	3,04155	38,01933	3,04155	38,01933
F_2	1,69275	21,15934	4,73429	59,17867
F_3	0,10162	1,27021	4,83591	60,44889
F_4	0,02170	0,27129	4,85761	60,72017

Tabela 6

Metoda największej wiarygodności

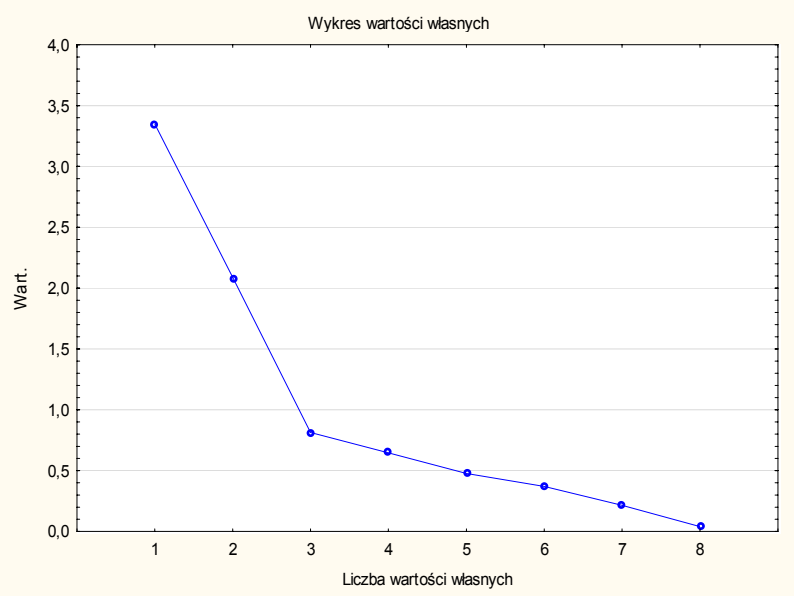
	Wartość własna	h_j (%)	Skumulowana wartość własna	H_k (%)
F_1	2,89747	36,21835	2,89747	36,21835
F_2	1,97842	24,73027	4,87589	60,94862
F_3	0,27838	3,47970	5,15427	64,42832
F_4	0,09061	1,13264	5,24488	65,56097

Tabela 7

Metoda centroidalna

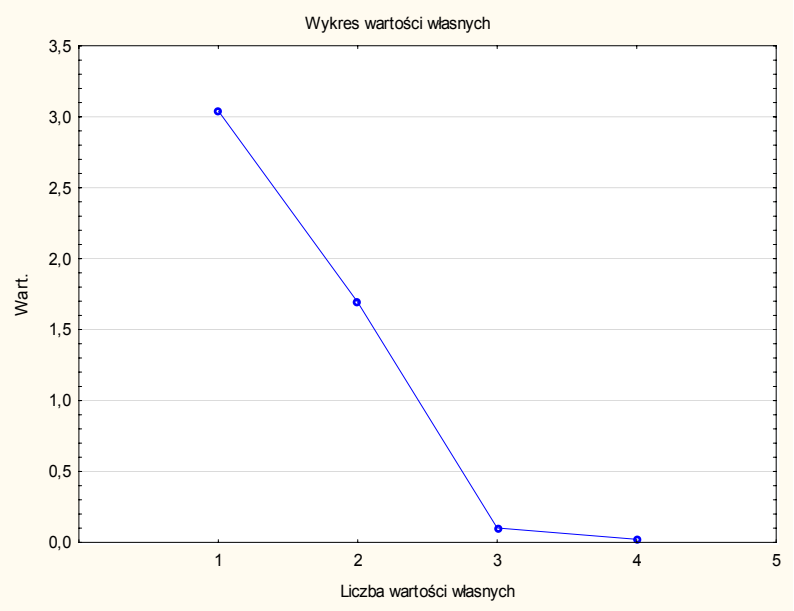
	Wartość własna	h_j (%)	Skumulowana wartość własna	H_k (%)
F_1	3,08785	38,59812	3,08785	38,59812
F_2	1,80870	22,60878	4,89655	61,20690
F_3	0,14521	1,81512	5,04176	63,02201
F_4	0,11341	1,41763	5,15517	64,43964

2. Wykresy osypiska:



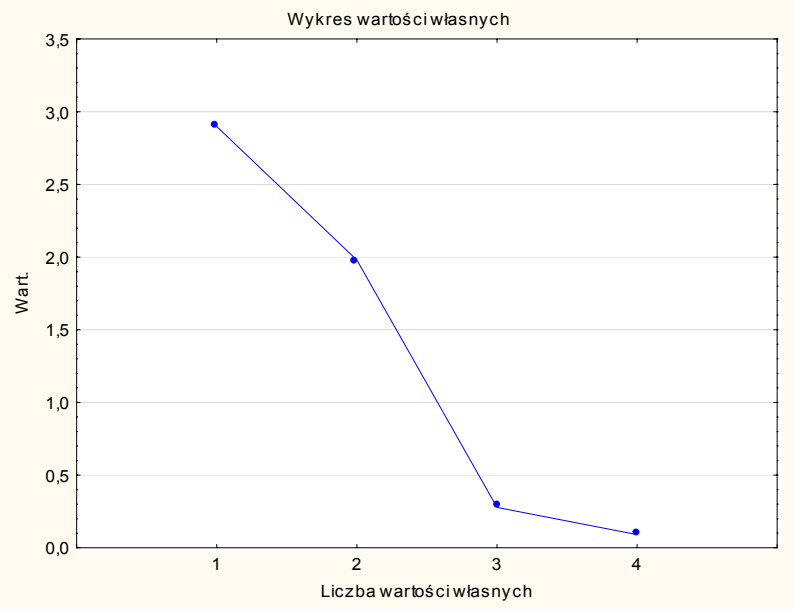
Rys. 2. Wykres osypiska – metoda składowych głównych

Źródło: Opracowanie własne z wykorzystaniem programu Statistica.



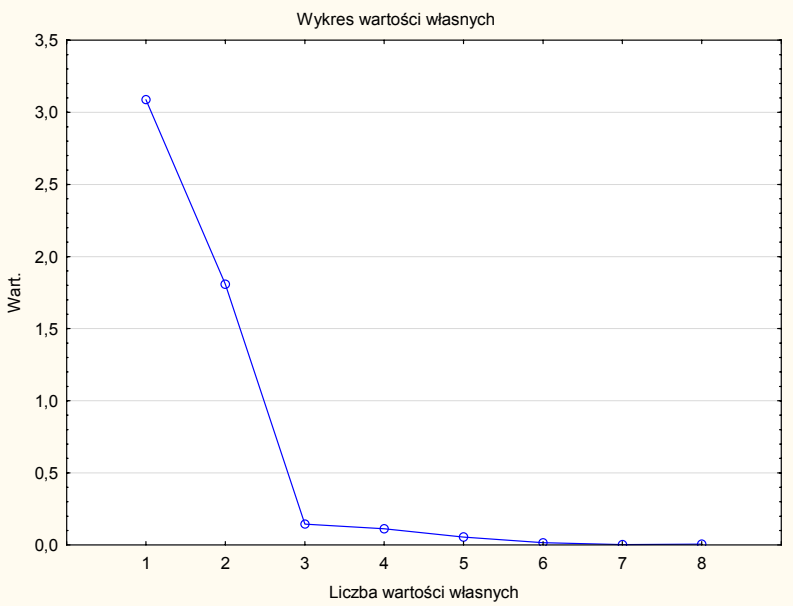
Rys. 3. Wykres osypiska – metoda głównego czynnika

Źródło: Opracowanie własne z wykorzystaniem programu Statistica.



Rys. 4. Wykres osypiska – metoda największej wiarygodności

Źródło: Opracowanie własne z wykorzystaniem programu Statistica.



Rys. 5. Wykres osypiska – metoda centroidalna

Źródło: Opracowanie własne z wykorzystaniem programu Statistica.

Zgodnie z wybranymi dwoma kryteriami – kryterium osypiska i kryterium wystarczającej proporcji, w którym dąży się do jak największego stopnia wyjaśnionej wariancji oryginalnych zmiennych – dokonano redukcji ośmiu zmiennych do trzech czynników w każdej z wybranych metod.

3. Tablice z wyodrębnionymi ładunkami czynnikowymi za pomocą wybranych metod z zastosowaną rotacją Varimax przedstawiono poniżej.

Tabela 8

Metoda składowych głównych

	F_1	F_2	F_3
BPP	**0,906088**	0,193881	0,210087
BZ	0,231644	0,105452	**0,941652**
BZW	**0,894938**	-0,128081	0,013673
BNPZ	**0,947446**	0,151940	0,174731
ZŚP	-0,028337	**0,846350**	0,118490
ZUP	0,162886	**0,727659**	-0,141502
ZCM	0,051718	**0,727890**	0,142561
PPP	0,192596	**0,838455**	0,165270

Tabela 9

Metoda głównego czynnika

	F_1	F_2	F_3
BPP	**0,937549**	0,182561	-0,150723
BZ	0,364813	0,191573	-0,131111
BZW	**0,785418**	-0,133661	0,233355
BNPZ	**0,971192**	0,138627	0,030187
ZŚP	0,028123	**0,783307**	0,021218
ZUP	0,127301	**0,571153**	0,044913
ZCM	0,103189	**0,618082**	0,023434
PPP	0,227712	**0,804953**	-0,095227

Tabela 10

Metoda największej wiarygodności

	F_1	F_2	F_3
BPP	**0,952829**	0,172299	-0,202474
BZ	0,367004	0,190854	-0,079508
BZW	**0,787298**	-0,134078	0,253331
BNPZ	**0,981287**	0,140274	0,064358
ZŚP	0,021719	**0,830838**	0,116879
ZUP	0,130949	**0,571697**	0,005766
ZCM	0,089348	**0,616609**	-0,033913
PPP	0,234874	**0,825477**	-0,181597

Tabela 11

Metoda centroidalna

	F_1	F_2	F_3
BPP	**0,932927**	0,186528	0,175063
BZ	0,361677	0,190384	0,147319
BZW	**0,785637**	-0,131191	-0,167420
BNPZ	**0,990931**	0,135145	-0,053094
ZŚP	0,025362	**0,811870**	-0,079666
ZUP	0,124214	**0,573591**	-0,115914
ZCM	0,106654	**0,616626**	0,006779
PPP	0,220190	**0,837739**	0,134804

Uzyskane wyniki wskazują, że najefektywniejszą metodą redukcji zmiennych w analizie czynnikowej jest metoda głównych składowych z zastosowaną

rotacją Varimax. Wybór trzech czynników w tej metodzie pozwolił na wyjaśnienie 78% całkowitej zmienności, co spełnia kryterium wystarczającej proporcji. Metoda ta w najlepszym stopniu przybliżyła wyniki analizy do tzw. prostej struktury, każda zmienna jest wysoko skorelowana tylko z jednym czynnikiem.

W pozostałych trzech metodach wybór trzeciego czynnika wydaje się zbędny. Niewiele on wnosi do wyjaśnienia całkowitej zmienności, która sięga mimo wszystko znacznie poniżej wymaganego poziomu 75%. Wyznaczone czynniki zachowują stosunkowo niedużą część informacji zawartych w zmiennych pierwotnych. Co więcej, zmienna BZ nie jest powiązana z żadnym czynnikiem, nawet dodanie czwartego czynnika nie zmieniłoby tej sytuacji.

Interpretując zatem wyniki analizy czynnikowej za pomocą metody składowych głównych, czynnik pierwszy wykazuje najwyższe ładunki dla zmiennych BPP, BZW oraz BNPZ, a więc jest związany głównie z bezrobociem. Czynnik drugi jest najwyżej skorelowany ze zmiennymi ZŚP, ZUP, ZCM oraz PPP, a więc jest związany z zatrudnieniem w warunkach zagrożenia i poszkodowaniem w wypadkach przy pracy, ogólnie dotyczy ciężkich warunków pracy. Czynnik trzeci, najsilniej związany ze zmienną BZ, również dotyczy bezrobotnych, ale konkretnie bezrobotnych zwolnionych. Można się zatem pokusić o następujące nazwy dla opisanych czynników:

– czynnik pierwszy – „Bezrobocie”,
– czynnik drugi – „Ciężkie warunki pracy”,
– czynnik trzeci – „Zwolnienie”.

Podsumowanie

Celem artykułu było porównanie efektywności analizy składowych głównych i analizy czynnikowej. Obie metody służą do redukcji zmiennych oraz do wyjaśniania istniejących korelacji między zmiennymi za pomocą kilku nieobserwowalnych i nieskorelowanych składowych głównych czy czynników. Do badań posłużyły dane z Rocznika Statystycznego Pracy 2010. Analizie poddano 311 powiatów Polski ze względu na osiem zmiennych, które w konsekwencji w analizie składowych głównych oraz analizie czynnikowej zostały zredukowane do trzech składowych i trzech czynników.

Analiza wybranego przykładu wykazała, iż wyniki otrzymane drogą analizy czynnikowej wykorzystującej metodę głównych składowych łatwiej poddają się interpretacji niż wyniki analizy składowych głównych. Wpływ na to ma niewątpliwie możliwość wykorzystania rotacji. W tym przypadku analiza czynnikowa okazała się efektywniejsza.

Literatura

1. Czyż T.: *Zastosowanie metody analizy czynnikowej do badania ekonomicznej struktury regionalnej Polski*. Wydawnictwo Polskiej Akademii Nauk, Wrocław 1971.
2. Frątczak E.: *Wielowymiarowa analiza statystyczna. Teoria – przykłady zastosowań z systemem SAS*. Szkoła Główna Handlowa, Warszawa 2009.
3. Grabiński T.: *Metody taksonometrii*. Akademia Ekonomiczna, Kraków 1992.
4. Krzyśko M.: *Wielowymiarowa analiza statystyczna*. Wydawnictwo Naukowe UAM, Poznań 2000.
5. Morrison D.F.: *Wielowymiarowa analiza statystyczna*. Państwowe Wydawnictwo Naukowe, Warszawa 1990.
6. Pluta W.: *Wielowymiarowa analiza porównawcza w modelowaniu ekonometrycznym*. Państwowe Wydawnictwo Naukowe, Warszawa 1986.
7. Pluta W.: *Wielowymiarowa analiza porównawcza w badaniach ekonomicznych*. Państwowe Wydawnictwo Ekonomiczne, Warszawa 1977.
8. Stanisz A.: *Przystępny kurs statystyki z zastosowaniem Statistica PL na przykładach z medycyny*. T. 3: *Analizy wielowymiarowe*. StatSoft, Kraków 2007.
9. Walesiak M., Gatnar E.: *Statystyczna analiza danych z wykorzystaniem programu R*. Wydawnictwo Naukowe PWN, Warszawa 2009.

COMPARATIVE ANALYSIS OF EFFECTIVENESS OF THE METHODS FOR REDUCTION OF VARIABLES – PRINCIPAL COMPONENT ANALYSIS AND FACTOR ANALYSIS

Summary

Principal component analysis and factor analysis are the two most popular methods that allow to bring a large number of studied variables to a much smaller number of mutually independent principal components or factors. New variables (principal components or factors) retain a relatively large part of the information contained in the original variables, while each of them is a carrier of other substantive content. Both of these methods of reduction of the variables are often used, because too many pending attributes increases the range of the difficulty of interpretation.

The main reason of undertaking the project is an attempt to show, that the abovementioned methods, although they are very similar, cannot be indentified. Despite the fact, that in both cases eigenvalues are calculated, factor loadings, etc., but still there are differences in the way of action, about which it must be remembered. So the usage of these names the variables are unacceptable.

The article consists of three parts. The first and second chapter are devoted, respectively, to the analysis of the principal components and factor analysis, where a short characterization of these methods had been made. In the third chapter, on the basis of an empirical example, we compared the effectiveness of the principal components analysis and factor analysis.